Spring 2021

# Regularized Deep Network Learning For Multi-Label Visual Recognition

Hao Guo

# Regularized Deep Network Learning for Multi-label Visual Recognition

by

Hao Guo

Bachelor of Science
Beijing Jiaotong University 2012

Master of Science
Beijing Jiaotong University 2015

---

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Computer Science

College of Engineering and Computing

University of South Carolina

2021

Accepted by:

Song Wang, Major Professor

Michael N. Huhns, Committee Member

Yan Tong, Committee Member

Ioannis Rekleitis, Committee Member

Xiaofeng Wang, Committee Member

Tracey L. Weldon, Interim Vice Provost and Dean of the Graduate School

# Acknowledgments

I would like to extend thanks to many people, from whom I received help during my Ph.D. pursuing.

First, I would like to express the deepest appreciation to my advisor, Prof. Song Wang, for the tremendous help and support he has provided for me. He has insightful and sharp sense of problems, which guides me through the difficulties in my research. He is patient and spends a lot of time to discuss with students to broaden our minds. He is rigorous about the work and is very careful of revising articles. I am so grateful to him for holding me to a high research standard and teaching me how to do research. Without his guidance and persistent help, this dissertation would not be possible.

I would like to thank my dissertation committee members, Prof. Michael Huhns, Prof. Yan Tong, Prof. Ioannis Rekleitis, and Prof. Xiaofeng Wang, for their insightful and constructive suggestions on my work. I want to thank them for their time. It is my great honor to have them serve on my dissertation committee.

I would also like to thank my colleagues: Xiaochuan Fan, Yuewei Lin, Shizhong Han, Zibo Meng, Hongkai Yu, Dazhou Guo, Kang Zheng, Yang Mi, Haozhou Yu, Yuhang Lu, Lan Fu, Zhenyao Wu, Xinyi Wu, James O'Reilly, Jie Cai, Zhiyuan Li and other fellow lab-mates. They give me a lot of help and encouragement.

Finally, I would like to thank my parents for their love and support during the past years. Their love kindles my enthusiasm during my difficult times. Their encouragement fuels up my study and research.

# Abstract

This dissertation is focused on the task of multi-label visual recognition, a fundamental task of computer vision. It aims to tell the presence of multiple visual classes from the input image, where the visual classes, such as objects, scenes, attributes, etc., are usually defined as image labels. Due to the prosperous deep networks, this task has been widely studied and significantly improved in recent years. However, it remains a challenging task due to appearance complexity of multiple visual contents co-occurring in one image. This research explores to regularize the deep network learning for multi-label visual recognition.

First, an attention concentration method is proposed to refine the deep network learning for human attribute recognition, i.e., a challenging instance of multi-label visual recognition. Here the visual attention of deep networks, in terms of attention maps, is an imitation of human attention in visual recognition. Derived by the deep network with only label-level supervision, attention maps interpretively highlight areas indicating the most relevant regions that contribute most to the final network prediction. Based on the observation that human attributes are usually depicted by local image regions, the added attention concentration enhances the deep network learning for human attribute recognition by forcing the recognition on compact attribute-relevant regions.

Second, inspired by the consistent relevance between a visual class and an image region, an attention consistency strategy is explored and enforced during deep network learning for human attribute recognition. Specifically, two kinds of attention consistency are studied in this dissertation, including the equivariance under spatial

transforms, such as flipping, scaling and rotation, and the invariance between different networks for recognizing the same attribute from the same image. These two kinds of attention consistency are formulated as a unified attention consistency loss and combined with the traditional classification loss for network learning. Experiments on public datasets verify its effectiveness by achieving new state-of-the-art performance for human attribute recognition.

Finally, to address the long-tailed category distribution of multi-label visual recognition, the collaborative learning between using uniform and re-balanced samplings is proposed for regularizing the network training. While the uniform sampling leads to relatively low performance on tail classes, re-balanced sampling can improve the performance on tail classes, but may also hurt the performance on head classes in network training due to label co-occurrence. This research proposes a new approach to train on both class-biased samplings in a collaborative way, resulting in performance improvement for both head and tail classes. Based on a two-branch network taking the uniform sampling and re-balanced sampling as the inputs, respectively, a cross-branch loss enforces consistency when the same input goes through the two branches. The experimental results demonstrate that the proposed method significantly outperforms existing state-of-the-art methods on long-tailed multi-label visual recognition.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

In the past decades, computer vision has been developed to one of the most popular areas among artificial intelligence researches. By classifying an image into different categories/classes, visual recognition [70] is an fundamental task of computer vision. Its goal is to tell whether an image contains certain objects, scenes, attributes, etc., each of which can be denoted as an image label. Based on the number of image labels associated with one image, we usually group the visual recognition as the single-label visual recognition and the multi-label visual recognition. For the traditional single-label visual recognition, there is only one category associated with an image, i.e., the existence of one category excludes the existence of other categories. As shown in Fig. 1.1(a), images are recognized as either "cat", "dog", "person" or "bicycle". In this case, the label annotation for each image is one-hot. However, in many applications, an image is usually associated with multiple (more than one) categories, as shown in Fig. 1.1(b). For example, the first image in Fig.1.1(b) contains both a dog and a cat and therefore is associated with both labels of "dog" and "cat". This dissertation is focused on the problem of multi-label visual recognition.

Multi-label visual recognition [149, 169] aims to classify an image into multiple classes/categories, denoted as multiple labels. Typical topics of multi-label visual recognition include human/pedestrian attribute recognition [4, 23, 41, 85, 90, 171], scene understanding [128], multi-object recognition [14], facial attribute recognition [46], etc. For example, for human attribute recognition (HAR), the goal is to identify the multiple labels that represent a set of pre-defined human attributes,


Figure 1.1 An illustration (a) single-label and (b) multi-label visual recognition. "1" indicates the presence and "0" represents absence of an image label. The four numbers below each image indicate the categories of "dog", "cat", "person" and "bicycle" sequentially.

e.g., "male", "sunglasses", "hat", "T-shirt" and "short", as shown in Fig. 1.2(a). As an important visual concept, human attributes are highly intuitive, semantic and informative to describe the appearance of a person. Accurate human attribute recognition can benefit a wide variety of computer vision applications, such as person re-identification [44, 137, 93], person retrieval [33], pedestrian detection [147], people search [150], fine-grained recognition [26], object categorization [74], object description [31], face verification [71] , and attribute-based classification [73], which is another kind of multi-label vision recognition. Figure 1.2(b) also shows an example of multi-object recognition, i.e., recognizing the existence of "person" and "car".

## 1.1 Challenges

Many advanced deep neural networks have been developed for enhancing the performance of multi-label visual recognition, such as exploring label dependencies [153, 154, 125, 174, 45, 88, 141, 7, 89, 173, 54, 87, 153, 178] and discovering the label-relevant regions [171, 84, 95, 97, 178, 124, 143, 142] or label-related contexts [90, 154]. How-

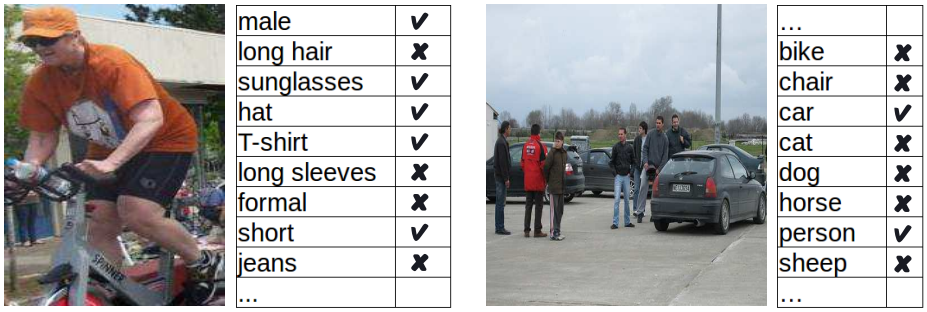


Figure 1.2 An illustration of multi-label visual recognition: (a) human attribute recognition; (b) multi-object classification.

ever, it remains as a very challenging task due to multiple image labels associated with one image, denoted as label co-occurrence. The label co-occurrence can lead to the challenges of label locality, label occlusion and label imbalance in multi-label visual recognition.

### 1.1.1 Label Locality

The label locality of multi-label visual recognition results from samples of multiple categories sharing the same image space. When an image is associated with multiple image labels, samples of each associated label should occupy certain image regions, as illustrated in Fig. 1.1(b). Given the image with a specific dimension, the sample resolution of a specific label in multi-label visual recognition could be suppressed by the number of co-occurred image labels. In this case, an image label is usually depicted by a local image region, defined as label locality in this dissertation. The remaining regions sometimes can be distractions for recognizing the specific image label. To be specific, with the human attribute recognition as the example, the attribute of "short" is only depicted by the local regions around upper legs of the person, while the attribute of "hat" exists at the head regions of the person, as shown in Fig. 1.2(a). Besides, even if the original image is in high resolution or high quality,

the image information associated to an attribute may be in low resolution and low quality in practice, e.g., the attribute "sunglasses" in Fig. 1.2(a). Thus, by limiting label-specific information, the label locality could increase the difficulty of recognizing the certain image labels/human attributes.

### 1.1.2 Label Occlusion

When samples of multiple categories exist in the same image, a sample of one category could occlude samples from other categories. For example, in human attribute recognition, the attributes "hat" and "long hair" are both recognized from the head regions of a person. As shown in Fig. 1.3, when an attribute occludes with another attribute, it makes the recognition of the occluded attribute more difficult.

Figure 1.3 An illustration of label occlusion in human attribute recognition of "hat" and "long hair".

### 1.1.3 Label Imbalance

Nowadays, deep neural networks are usually trained to address the multi-label visual recognition, which requires a large-scale dataset containing images with multiple labels annotated. Practically, efforts required for including images of different categories are different. Thus, image datasets for multi-label visual recognition are usually imbalanced, i.e., some categories have more samples than other categories. Deep network training from the imbalanced data normally biases towards the categories with

more training samples. For single-label visual recognition, this issue can be addressed by artificially balancing the categories, such as including equal number of images for each category in ImageNet [22]. But achieving category balancing is not an trivial work for multi-label visual recognition, since adding or removing an image may include or delete samples of multiple categories, because of label co-occurrence. Besides, for multi-label visual recognition, recognizing the presence of each image label is regarded as equally important. Thus, the imbalanced label distribution is also a challenge for multi-label visual recognition.

## 1.2 Research Scope

To address the multi-label visual recognition, this dissertation explores and studies the label locality and label imbalance for deep network regularization. Specifically, the label locality is studied by regularizing the deep network attention, while the label imbalance is addressed by handling the long-tailed issue in multi-label visual recognition.

### 1.2.1 Label Locality: Deep Network Attention Regularization

To explore label locality for multi-label visual recognition, we address the specific task of human attribute recognition, since the recognition of human attributes is usually determined by certain regions of the image. Based on this attribute locality, the attribute-region relevance plays an important role in human attribute recognition. This leads to one of the most important properties of human attributes, which is denoted as *local spatiality* in this dissertation, i.e., an attribute is usually related to particular human-body parts, local image regions, or certain contexts [171, 84, 95, 97, 178, 124, 143, 142]. Such attribute-region relevance is frequently reflected by the attention mechanism. Prior researches in cognitive science [102, 75] and neuroscience [24] show that our human vision actually recognizes an attribute

by discovering and focusing visual attention on such local discriminative regions. By simulating this attention in human vision [69, 27, 148, 68, 18], many deep networks [175, 127, 3] have been designed to generate attention maps by identifying the local image regions that contribute most to the final recognition. They explain why the deep network makes certain prediction, no matter if the prediction is correct or not, i.e., the highlighted areas are discovered by the deep network as the attribute-relevant regions. In most cases, such attention maps are computationally estimated as an intermediate result (or a byproduct) of the model prediction with only image-level supervision, and have been widely used for network interpretation. In this dissertation, the interpretive attention maps of recognizing human attributes are leveraged for further enhancing the performance of human attribute recognition.

Ideally, according to the local spatiality of human attributes, if a deep network is well trained and completely robust for recognizing an attribute, it shall precisely focus attention on attribute-relevant regions, e.g., face regions for attribute "sunglasses". However, in practice, attention maps from existing deep networks can not always highlight regions semantically relevant to the corresponding attributes. As shown in Fig. 1.4, when a ResNet50 [49] is used to recognize the attributes (a) "sunglasses" and (b) "male", the estimated attention maps highlight irrelevant regions to these attributes, respectively, and correspond to incorrect predictions. Thus, a basic assumption can be made that the correctness of attention maps reflects the performance of the trained deep network for attribute recognition. Fig. 1.5 shows an example of attention map changing during a ResNet50 training for human attribute recognition. We can see that, with more training iterations, the attention maps of attribute "sunglasses" on these two images are getting more focused on the desired face regions, i.e., the attribute-relevant regions, and meanwhile, the network is getting better trained by outputting more accurate prediction scores, i.e., higher score for the attribute presence in the top image and lower score for the attribute absence in the bottom


Figure 1.4 Illustration of incorrect attention maps for recognizing attributes of (a) sunglasses and (b) male.

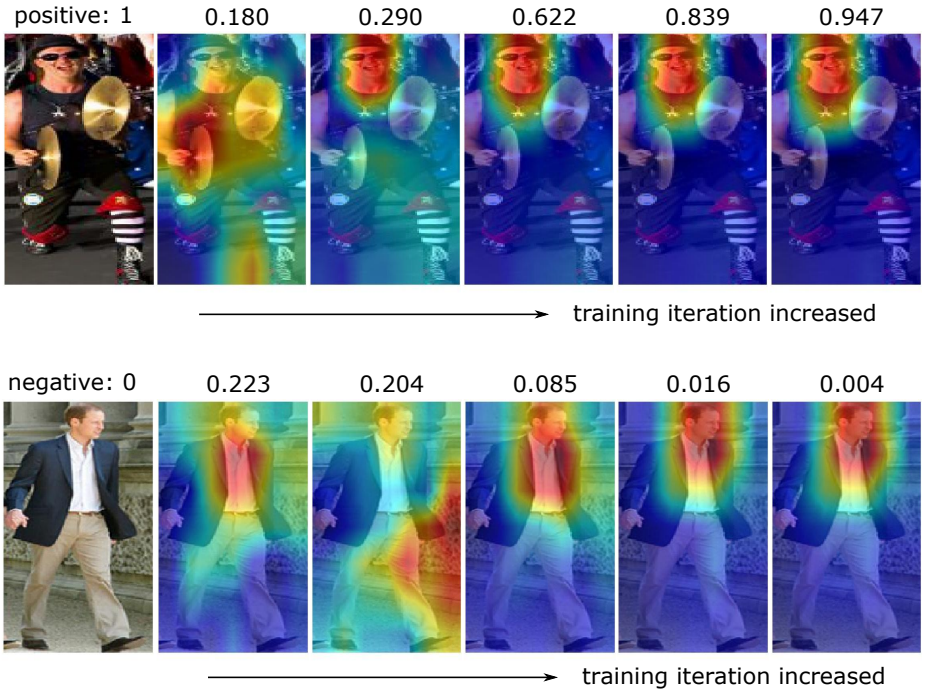


Figure 1.5 Attention maps for attribute "sunglasses" in different iterations of a deep network (ResNet50) training, where face is the desired attribute-relevant region. The number above each attention map represents the predicted presence score (in [0, 1]) in the corresponding iteration.

image. Motivated by this, this dissertation proposes to incorporate the plausibility of attention maps into the network to regularize the network training for improving human attribute recognition. This essentially utilizes the local spatiality of human attributes and addresses the attribute locality for human attribute recognition.

The straightforward approach to achieve this goal is to impose implicit supervision of attribute-relevant regions in deep network learning, which requires the pixel-level ground-truth of attention maps, which are similar to the ground truth used for semantic segmentation [2]. However, annotating such pixel-level ground truth for large-scale training images is infeasible due to cognitive ambiguity and intensive labor involved in manual annotations:

- cognitive ambiguity – attribute-relevant regions are not well defined: (a) It is not a trivial work to identify regions relevant to certain abstract attributes, such as "Age Between 18 and 60"; (b) Different from discrete values in pixel-level ground truth for semantic segmentation, pixel values of attention map annotations are continuous, leading to the difficulty of quantifying the importance of each pixel for recognizing an attribute.

- intensive labor – multiple attention maps need to be annotated for each image: in the same image, different attributes have different attention maps, which heavily increases the difficulty and workload of manual annotation.

Therefore, to achieve fully supervision on attention maps of attributes, a comprehensive group study is required to define the relevant regions for each attribute and the annotation consumes a lot of labor and time. Thus, this dissertation explores and designs methods to improve the plausibility of attention maps for regularizing the deep network learning without requiring ground-truth attention maps.

### 1.2.2 Label Imbalance: Distribution Balancing on Long-tailed Multi-label Image Data

Existing state-of-the-art methods for multi-label visual recognition are usually based on deep networks, which are data-driven. The quality of training data decides the robustness of the trained networks. Therefore, many benchmarks, e.g., Ima-

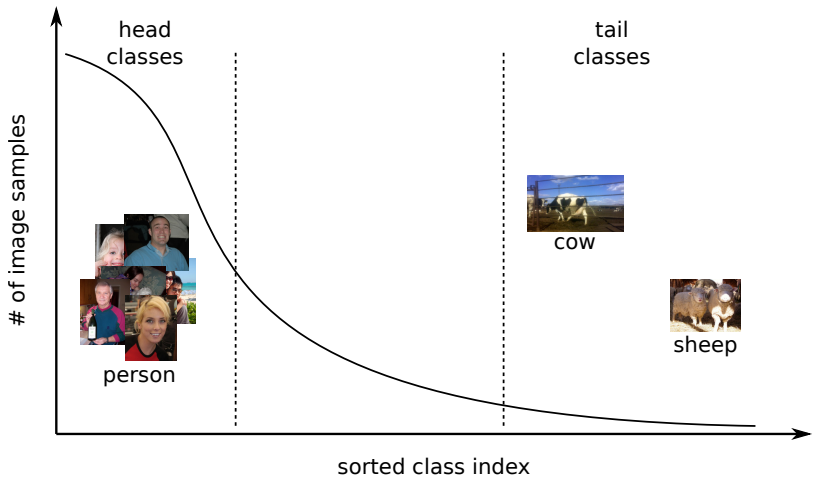


Figure 1.6 An illustration of long-tailed distribution in a set of images. In this case, "person" is one of the head classes, while "cow" and "sheep" are two of tail classes. The class indexes are sorted according to the number of images, in the descending order.

geNet [22, 121], MS-COCO [92], are usually constructed by artificially balancing the number of images for each class. However, in practice, the numbers of images for different classes are more likely imbalanced. As discussed in some single-label visual recognition works [98, 8, 19, 157, 94, 61, 65, 176], the numbers of training images for different classes may exhibit a long-tailed distribution in terms of classes (image labels), as shown in Fig. 1.6. The *head classes* have many image samples, while *tail classes* have very few image samples in training data. Direct training on such data (with uniform sampling) usually produces relatively low performance on the tail classes. To address this issue, many re-balancing methods are proposed for single-label visual recognition. Due to label co-occurrence, i.e., multiple image labels associated with one image, the long-tailed issue in multi-label visual recognition [166] is more challenging, and has not been well explored.

For example, re-balanced data sampling [10, 129, 6, 47] is a proven effective approach for addressing the long-tailed visual recognition. It achieves class-wise balance by either down-sampling the head-class data or up-sampling the tail-class data,

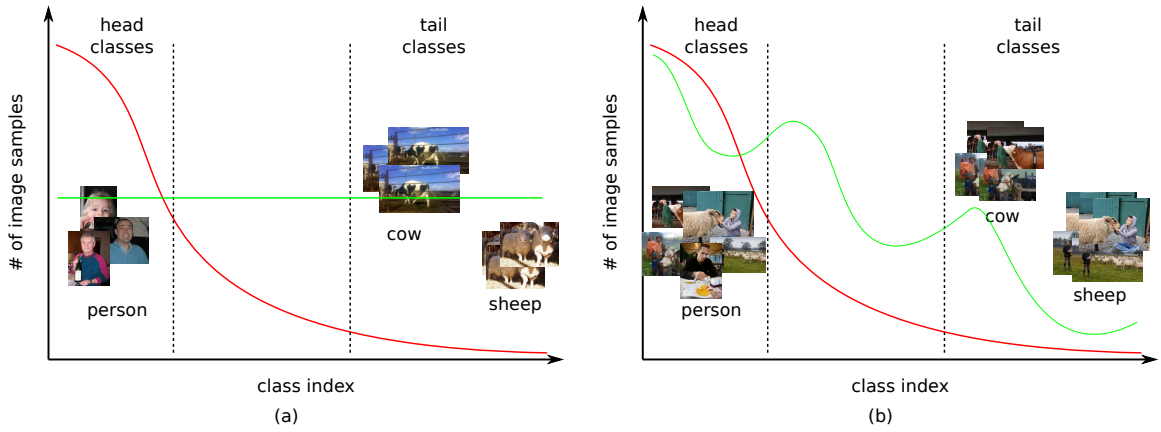


Figure 1.7 The illustration of using re-balanced sampling to address the long-tailed issue in (a) single-label visual recognition and (b) multi-label visual recognition. Red curves represent the original long-tailed distribution, while green curves illustrate the re-balanced distributions.

as shown in Fig. 1.7(a), leading to the performance improvement on tail classes. However, the re-balanced sampling can not directly achieve class-wise balance for multi-label visual recognition. Since each image may be associated with multiple classes/labels, down-sampling an image of a head class or up-sampling an image of a tail class may simultaneously decrease the images for tail classes or increase the images for head classes, respectively. Instead of achieving a class-wise balanced distribution, the re-balanced sampling intends to yield another imbalanced distribution, as shown in Fig. 1.7(b). Thus, while the re-balanced sampling can improve the recognition performance of tail classes, it may simultaneously suppress the performance of some head classes due to label co-occurrence in multi-label visual recognition [166]. Since performance of different classes, either head or tail ones, is usually considered to be equally important in multi-label visual recognition, this dissertation develops a new method that can combine different data samplings for improving the performance of both head and tail classes.

## 1.3 Proposed Approaches

According to the above discussions, this dissertation proposes visual attention concentration and visual attention consistency to integrate the label locality for multi-label visual recognition, i.e., attribute locality in human attribute recognition, and collaborative learning on biased distributions to address the label imbalance, i.e., long-tailed distribution, in multi-label visual recognition.

### 1.3.1 Visual Attention Concentration

Considering the relevance between attributes and human body parts, prior part-based methods extract features at human body parts corresponding to each human attribute and the part-based features are fed to classifiers individually or together for recognizing human attributes. This verifies that human attribute recognition can be achieved on certain image regions. However, the performance of these methods is highly dependent on the accuracy of body-part detection, which is a well known challenging problem in computer vision. They also require predefined correlation between attributes and body parts for attribute recognition, while this correlation sometimes is not well defined for certain attributes, as discussed in Section 1.2.1. Therefore, this study proposes an indirect method to enforce the deep networks to recognize human attributes from attribute-relevant regions.

When the ordinary deep network learning minimizes the image classification loss for attribute recognition, the attention maps estimated by class activation mapping [175] highlight image areas regarded as attribute-relevant regions in the network's opinion. As the deep network could not be well trained with limited training data, the highlighted regions may be actually irrelevant to certain attributes. Then, this study introduces an adversarial component to enhance the confidence of the highlighted regions by enforcing their concentration. For the highlighted regions relevant to attributes, enforcing their concentration emphasizes these regions for attribute

recognition, leading to refined recognition results. For the highlighted regions irrelevant to attributes, enforcing their concentration suppresses the regions really relevant to attribute recognition, which confronts with minimizing the classification loss for attribute learning, and propels the network to discover another area to highlight, i.e., for attribute recognition. Thus, the newly introduced component and the ordinary classification loss adversarially learn the deep network for better human attribute recognition. To achieve this goal, this study proposes an exponential loss applied to each attention maps for each attribute recognition to emphasize the highlighted regions and suppress the remaining regions.

The proposed method regularizes the deep network learning to focus attention on a single compact image region for each attribute recognition, which does not requires the body-part detection and the predefined correlation between attributes and body parts. Experimental results on two public datasets and two deep networks verify its effectiveness, by outperforming the results from part-based methods.

### 1.3.2 Visual Attention Consistency

Based on the assumption that more plausible attention maps indicate better networks in Fig. 1.5, this study proposes a new approach for human attribute recognition by exploring and enforcing attention consistency for network regularization. Given an attribute in an image, the attribute-region relevance is important for the attribute recognition and should be constant if the perception of the image is not changed. As a reflection of attribute-region relevance in deep networks, interpretive attention maps should also be consistent under certain circumstances. Thus, this study proposes to enforce attention consistency for network learning for attribute recognition. Specifically, two kinds of attention consistency are explored. One kind of consistency enforces the equivariance of the attention map when the input image undergoes certain spatial transforms, such as scaling, rotation and flipping. The other kind of

the consistency is enforced between the attention maps derived from two different networks when both of them are trained for recognizing the same attribute from the same image. These two kinds of consistency are formulated as new loss functions and combined with the traditional classification loss for network training. Experiments on three datasets of human attribute recognition verify the effectiveness of the proposed method by achieving new state-of-the-art performance.

### 1.3.3 Collaborative Learning on Biased Distributions

To address the long-tailed issue of multi-label visual recognition, this study proposes a new method to train a network from differently biased distributions in a collaborative way. Given the long-tailed distributed training data for multi-label visual recognition, training with the uniform sampling emphasizes the recognition of head classes, while training with the re-balanced sampling emphasizes the tail classes. Meanwhile, recognition of tail classes and head classes are underrated by the uniform sampling and the re-balanced sampling, respectively. Each of these two samplings make the network training from biased distributions. Therefore, this study designs a visual recognition network with two branches to leverage both samplings. One branch takes the uniform sampling from long-tailed training data as input, while the other branch takes the re-balanced sampling from long-tailed training data as the input. For each branch, the binary-cross-entropy-based classification loss with learnable logit compensation is conducted for the visual recognition of each image label. A new cross-branch loss is defined to enforce the consistency when the same input image goes through the two branches. As the two branches emphasize the recognition of head classes and tail classes, respectively, this consistency makes two branches learn from each other collaboratively. The collaborative training attempts to compromise between two branches and lead to an effect equivalent to learning from a more balanced implicit distribution somewhere between the two biased dis-

tributions serving as the inputs of two branches, respectively. Extensive experiments are conducted on two public datasets. The results show that the proposed method significantly outperforms previous state-of-the-art methods on long-tailed multi-label visual recognition.

## 1.4 Structure of the Dissertation

The remainder of this dissertation is organized as follows. Chapter 2 introduces the deep neural networks and the visual attention mechanism as the basis of this dissertation. In Chapter 3, a literature review for related works is conducted. Chapter 4 elaborates on the proposed method of enforcing visual attention concentration for human attribute recognition. Chapter 5 explores and enforces visual attention consistency for human attribute recognition. Chapter 6 addresses the long-tailed issue of multi-label visual recognition with collaborative learning. Finally, Chapter 7 concludes the dissertation and outlooks the future work.

# Chapter 2

# Background

## 2.1 Brief History of Deep Neural Networks

Deep neural networks are responsible for some greatest advances in computer vision field in recent years [1]. As a specific instance of deep neural network, the Convolutional Neural Networks (CNNs) originate from the concept of receptive field from study on cat's visual cortexes [58] by two Nobel Prize winners, David H. Hubel and Torsten N. Wiesel in 1950s and 1960s. Inspired by the receptive field, the Neocognitron [35] was introduced in 1980 and could be regarded as the first consideration of CNN architecture. It introduced two basic types of layers: convolutional layers and down-sampling layers.

In 1986, the back-propagation algorithm was proposed [118, 119, 120], which has been proved to be very effective and is standard in most CNNs today. Then, a great breakthrough of the CNN was made in 1989, the backpropagation was used to learn the convolutional kernel coefficients directly from images of hand-written digits [78] by Yann LeCun et al. This work was further improved in [79]. It is the first real CNN architecture and initiates the applications of CNN models in computer vision tasks, such as face detection [116, 117], face recognition [76], and character recognition [133].

As the LeNet-5 [80], which is a pioneering 7-level convolutional network, was proposed in 1998, the CNN architecture was basically settled. LeNet-5 was applied by several banks to recognize hand-written numbers digitized in $32 \times 32$ pixel images.

[1] https://amturing.acm.org

To process higher resolution images, larger and more layers of convolutional neural networks are required, which is constrained by the availability of computing resources. Therefore, the development of convolutional neural networks was almost frozen at the beginning of 2000s.

In 2004, the Graphics Processing Units (GPUs) were used for training neural networks [106], which proved that GPUs can greatly accelerate the standard neural networks. In 2006, the GPUs were applied to CNNs [11]. Then, researchers started to use GPUs to accelerate neural network computing [16, 15] extensively.

With the advances in computation power benefiting from GPUs, the prosperous studies on CNNs are actually started from the AlexNet [70], when the massive datasets, such as ImageNet [121], became available. In 2012, when AlexNet, a convolutional neural network consisting of convolutional layers, pooling layers, activation layers (ReLU) and fully connected layers, was proposed, it achieved a top-5 error of 15.3% on ImageNet 2012 Challenge, which is more than 10 percentage points better than that of the runner up. After this, explorations on CNNs sprung up all over the computer society. According to The Economist, "Suddenly people started to pay attention, not just within the AI community but across the technology industry as a whole."

Then, a number of outstanding architectures are developed to make CNNs more robust for a wide variety of tasks, such as VGGNet [135], GoogLeNet [140], ResNet [49], DenseNet [57] for recognition, R-CNN [37], Fast R-CNN [36], Faster R-CNN [113], Mask R-CNN [48], SSD [96], YOLO [111] for detection, FCN [99], SegNet [2] for segmentation, and so on, as shown in Fig. 2.1. In summary, convolutional neural networks now have been an industry standard in computer vision.

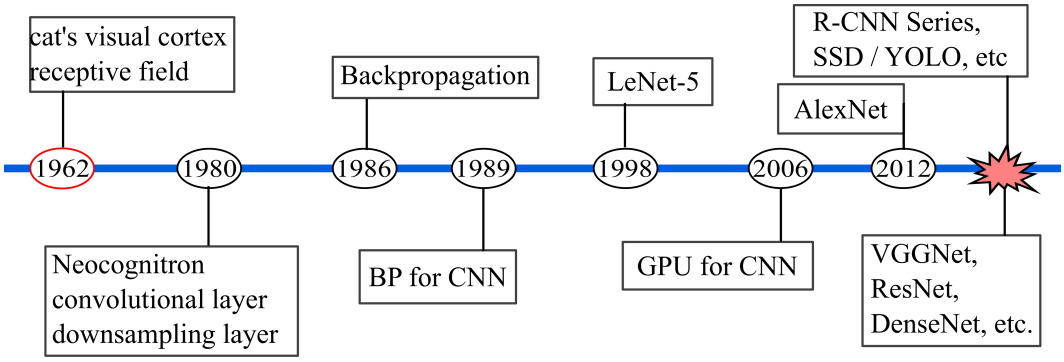


Figure 2.1 Some of the landmark events in the CNN history.

## 2.2 Architecture of CNNs

The deep convolutional neural networks are specific instances of Neural Networks (NNs), which simulates human brains. NNs are usually made up of relatively simple computing elements called "neurons", which loosely resemble the neurons in human brains. The neurons influence one another via weighted connections. The training of a neural network learns the weights on the connections, which changes the computation performed by the neural network. Similar to the ordinary neural networks, convolutional neural networks also consist of neurons that have learnable weights, i.e., convolutional kernels. Each neuron receives some inputs, performs convolutional computation and is followed by certain non-linearity operation.

In general, a typical CNN architecture is built from a stack of layers, each of which take the output of the prior layer and estimate output as the input of the successive layer, except the beginning input layer and the final output layer [86, 66]. There are mainly three types of layers to build CNN architectures: Convolutional Layer, Pooling Layer, and Fully Connected Layer.

### 2.2.1 Convolutional Layer

The convolutional layer is the core component to build CNN architectures. It consists of a certain number of convolutional kernels, a.k.a filters, and is usually followed by a non-linear function, e.g., Sigmoid, Rectified Linear Unit (ReLU) [104], to bring in non-linearity. The learnable weights and biases of all filters compose the layer parameters. Given an input image as the input layer, a convolutional layer transforms the input with dimension of channel, height and width to a volume of activation maps (or feature maps) by conducting convolution on each pixel of the image based on each convolutional filter in this layer. The successive convolutional layers take the volume from the prior layer and conduct convolutional operations repeatedly. Fig. 2.2 shows an illustration of the convolution operations based on the filters in a convolutional layer. Suppose the convolutional layer takes the input in dimension of $c_1 \times h_1 \times w_1$, and has $c_2$ convolutional kernels with kernel size as $k \times k$. Then, each kernel has parameters in dimension of $c_2 \times k \times k$. The convolution operation between the input and a convolutional kernel yields a feature map in dimension of $h_2 \times w_2$. Thus, $c_2$ kernels produce the output volume in dimension of $c_2 \times h_2 \times w_2$. Here

$$\begin{cases} h_2 = \lceil \frac{h_1 - k + 2p}{s} \rceil + 1, \\ w_2 = \lceil \frac{w_1 - k + 2p}{s} \rceil + 1, \end{cases} \tag{2.1}$$

where $p$ is the spatial padding outside of the boundary of the input, $s$ is the stride when sliding the same convolutional kernel to location of the input.

The motivation behind the filters are biologically-inspired by *receptive field* from Hubel and Wiesel's work on the cat's visual cortex [58, 59]. This study reveals that the animal visual cortex contains neurons, which are sensitive to small-sub-regions of visual field, called a receptive field. Two of the core properties of the convolutional layer are local connectivity (a.k.a sparse connectivity) and shared weights.

- *Local Connectivity:* Instead of connecting each neuron to all regions in the input volume as in the ordinary Neural Network, the convolutional layers connect each

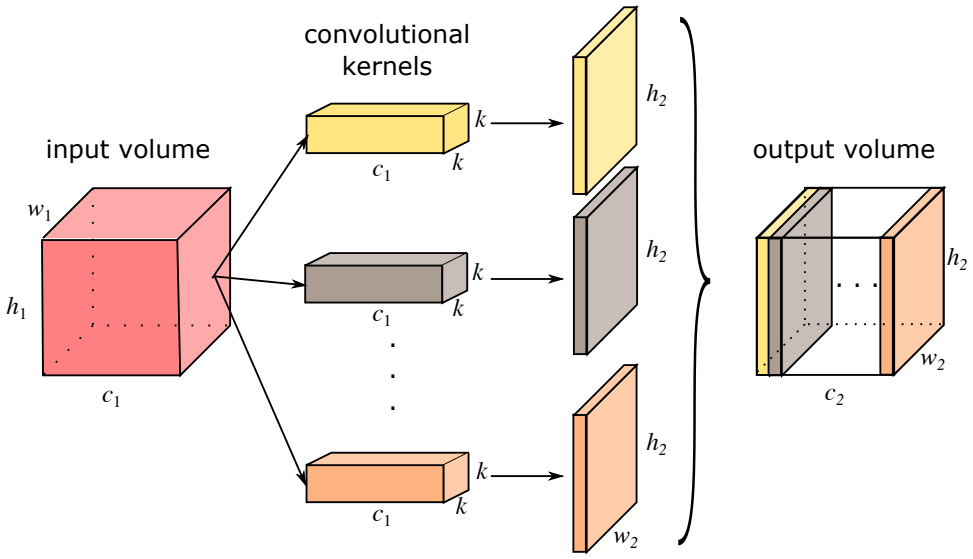


Figure 2.2 Each convolutional kernel (filter) convolves the input volume across the height and width, with extending through its full depth.

neuron only to a local region (kernel size) of the input volume, i.e., the receptive field of convolutional kernels. The local connectivity can obviously reduce the number of parameters to learn.

- *Shared Weights:* The strategy of shared weights in convolutional layers is also used to reduce the number of parameters. Intuitively, if a filter computes useful activation at a particular position of the input volume, it should also computes useful activation at a different position. Therefore, each filter filters the entire input volume in a sliding manner to produce a single depth slice of the output volume.

Generally, a non-linear activation function is appended to the convolutional layer to impose non-linearity to the network. It is applied individually to each element of the output volume of the convolutional layer. The dimension of feature maps do not change after activation functions. Several activation functions have been explored, such as Rectified Linear Unit (ReLU) [104], Sigmoid, TanH, etc. The most frequently

used one in today's CNN architectures is the ReLU function:

$$f(x) = \begin{cases} 0 & \text{for } x < 0, \\ x & \text{for } x \geq 0, \end{cases} \tag{2.2}$$

where $x$ is the input of the activation function $f$.

### 2.2.2 Pooling Layer

A Pooling Layer is used to down-sample the feature maps. The most common pooling layer is the MAX Pooling with a kernel size $2 \times 2$ and a stride 2. The maximum of 4 values from a $2 \times 2$ region on each depth of activation maps, is selected to represent the activation of this region on this depth slice. A simple illustration is shown in Fig. 2.3. There are also other pooling layers available, such as average pooling using the average of a region as representation, stochastic pooling randomly selecting a value in the region as its representation, etc.

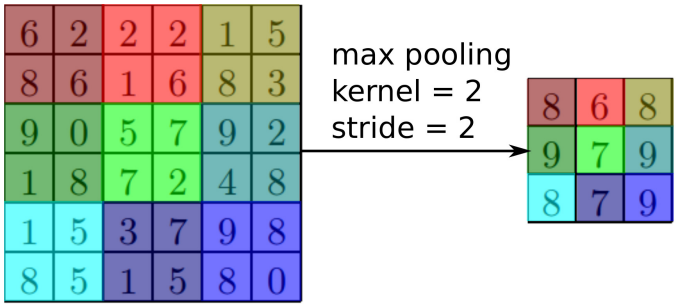


Figure 2.3 An illustration of max pooling applied on a single depth slice of feature maps.

### 2.2.3 Fully Connected Layer

The Fully Connected Layer is a linear layer widely used in ordinary neural networks. Each neuron in a fully connected layer is connected to all neurons of its prior layer. The computation of features from fully connected layers can be formulated as a matrix

multiplication with a bias offset:

$$\mathbf{Y} = \mathbf{W}^T\mathbf{X} + \mathbf{b}, \tag{2.3}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are the input and output of the fully connected layer, while $\mathbf{W}$ and $\mathbf{b}$ are the weights and biases of the fully connected layer. Actually, the fully connected layers now may not be mandatory to construct CNN architectures for certain computer vision tasks, e.g., FCN [99] for semantic segmentation.

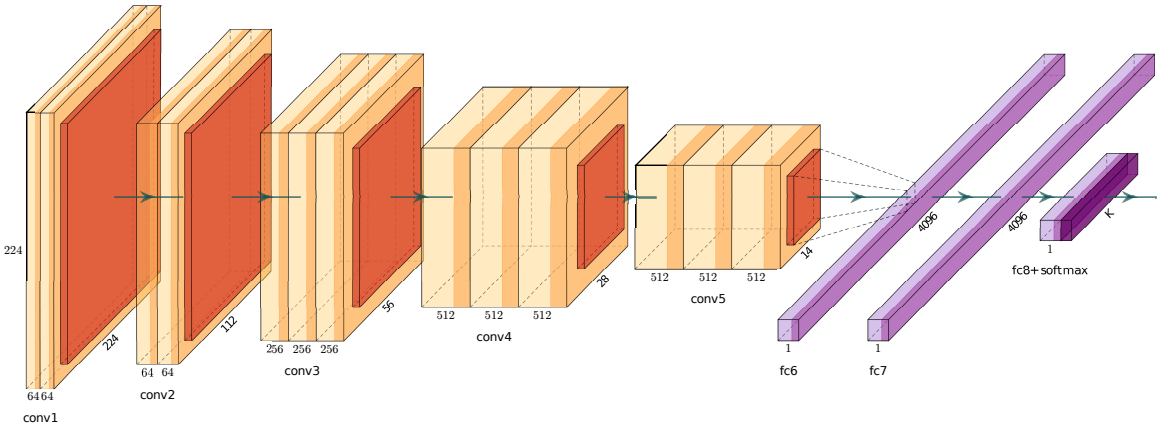


Figure 2.4 An illustration of a typical CNN architecture – VGG16.

Finally, an integrated CNN is built by stacking these layers together. For example, the classical VGG [135] is shown in Fig. 2.4 (drawn from PlotNeuralNet [2]). In this figure, "conv1", "conv2", ..., "conv4" are convolutional modules, each of which consists of a stack of convolutional layers and activation layers. The light yellow blocks represent the feature maps from convolutional layers. After activation function, i.e., ReLU, the block color changes to orange. The red blocks at the end of each convolutional module represent the down-sampled feature maps by pooling layers. Similarly, the light and dark purple blocks of the fully connected layers indicate the features before and after the activation function, respectively. Besides, "softmax" is a nor-

[2]https://github.com/HarisIqbal88/PlotNeuralNet

malization function to convert the network output to the final prediction. Numbers around blocks specify the dimension of feature maps/vectors at different stages.

## 2.3 Deep Network for Multi-label Visual Recognition

In this study, training deep networks for multi-label visual recognition is formulated as a problem of multiple binary image classification, as shown in Fig. 2.5. Given an image $\mathbf{x} \in \mathbb{X}$, the task aims to predict the presence of each image label, e.g., objects, human attributes, etc. The ground-truth for the image are $\mathbf{y} \in \mathbb{Y}$, with $\mathbf{y} = \{y_1, y_2, ..., y_K\}$ where $y_j = 1$ if image label $j$ is present in the image and $y_j = 0$ otherwise. $K$ is the number of considered image labels. $\mathbb{X}$ is the set of $N$ training images and $\mathbb{Y}$ is their corresponding set of ground-truth annotations.

For classical CNN architectures, such as AlexNet [70] and VGG16 [135], a CAM structure, which contains a Global Average Pooling (GAP) and an FC layer after the last convolutional layer of the CNN (denoted as GAP-FC), replaces the original stacked FC layers as shown in Fig. 2.4, so that CAM can be used for estimating the attention maps in network inference, which will be discussed in the following Section 2.4. Thus, in this study, these networks are denoted as AlexNet-CAM and VGG16-CAM, respectively. Other classical CNN architectures, such as ResNet [49], DenseNet [57], are designed with GAP-FC structure for producing the final results for visual recognition.

The input image $\mathbf{x}$ first goes through a sequence of convolutional layers, followed by certain activation functions and pooling layers. Convolutional feature maps from the last convolutional layer is fed to the global average pooling for feature aggregation. Then, an FC layer with multiple sets of linear weights and biases take the aggregated feature as input and output the final prediction for the presence of each image label. Usually, the binary cross entropy loss is used as the classification loss to learn the network for each image label recognition. For example, the cross-entropy-based

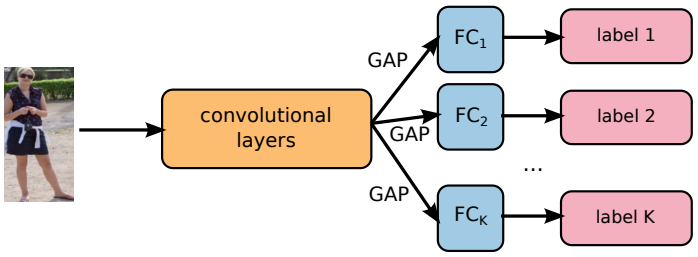


Figure 2.5 The illustration of CNNs with CAM structure, i.e., GAP-FC, being used for recognizing multiple image labels.

classification loss is widely used for human attribute recognition [83, 85, 97, 42, 141]. Suppose $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_K] \in \mathbb{R}^K$ denote the output of the deep network with $\mathbf{x}$ as the input. The classification loss would be computed as

$$\begin{aligned}\mathcal{L}_{cls}(\hat{\mathbf{y}}, \mathbf{y}) = -\frac{1}{K}\sum_{j=1}^{K}\omega_j \left(y_j \log(\varsigma(\hat{y}_j))\right. \\ \left. +(1-y_j)\log(1-\varsigma(\hat{y}_j))\right),\end{aligned} \tag{2.4}$$

where

$$\varsigma(\hat{y}_j) = 1/(1+e^{-\hat{y}_j}), \tag{2.5}$$

and $\hat{y}_j \in \mathbb{R}$ indicates the predicted score for image label $j$ being present in image $\mathbf{x}$. $\omega_j$ is used for weighting the loss from image label $j$ to alleviate the imbalance among different labels. Two weighting strategies are considered in the network training. The exponential strategy [83] produces relatively smooth attribute weights:

$$\omega_j^e = \begin{cases} e^{1-\rho_j} & \text{if } y_j = 1, \\ e^{\rho_j} & \text{if } y_j = 0, \end{cases} \tag{2.6}$$

where $\rho_j$ is the ratio of positive samples for image label $j$. The square root strategy [141] heavily emphasizes the attributes with rare positive samples:

$$\omega_j^s = \begin{cases} \sqrt{1/2\rho_j} & \text{if } y_j = 1, \\ \sqrt{1/2(1-\rho_j)} & \text{if } y_j = 0. \end{cases} \tag{2.7}$$

## 2.4 Class Activation Mapping for Deep Network Attention

Due to its differentiable and efficient computation, Class Activation Mapping (CAM) [175] is used for estimating visual attention of deep networks in this study. As the input image $\mathbf{x}$ is fed to the deep network (denoted as $f$) for multi-label recognition, a set of convolutional feature maps $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ is obtained from the last convolutional layer, where $C$, $H$ and $W$ are the channel, height and width of the feature maps, respectively. With the Global Average Pooling (GAP), the feature maps can be aggregated to a feature vector $\mathbf{f} \in \mathbb{R}^C$, which is passed to the linear layer, i.e., FC layer, for image label attribute recognition. The parameters of the linear layer for recognizing multiple image labels consist of linear weights $\mathbf{W} \in \mathbb{R}^{K \times C}$ and bias $\mathbf{b} \in \mathbb{R}^K$. The prediction for the presence of image label $j$ in image $\mathbf{x}$ can be written as

$$\hat{y}_j = \mathbf{w}_j \mathbf{f} + b_j, \tag{2.8}$$

where $\mathbf{w}_j \in \mathbb{R}^C$ is the $j$-th row of $\mathbf{W}$ and $b_j$ is the $j$-th element of $\mathbf{b}$, and they represent the linear weights and bias for recognizing image label $j$, respectively.

As each value in the feature vector $\mathbf{f}$ is aggregated from a channel (a visual pattern) of the feature maps, the learned linear weights $\mathbf{w}_j$ specify the importance of each visual pattern for recognizing the image label $j$. Therefore, as shown in Fig. 2.6, CAM directly maps the linear weights to the channels of feature maps $\mathbf{F}$ to estimate the $H \times W$-dimensional attention map

$$h(\mathbf{x}, j, f) = \sum_{c=1}^{C} w_{jc} \mathbf{F}_c \tag{2.9}$$

for label-$j$'s presence in image $\mathbf{x}$, where $w_{jc}$ is the $c$-th value of $\mathbf{w}_j$, and $\mathbf{F}_c \in \mathbb{R}^{H \times W}$ is the $c$-th channel of the feature maps $\mathbf{F}$. Note that $\mathbf{w}_j$ and $\mathbf{F}$ are derived from $f$ and $\mathbf{x}$, respectively. Therefore, CAM is denoted as a function of both the input image and the network in Eq. (2.9). Using bi-linear interpolation, the attention map can be up-sampled to the input image size to indicate pixel-level evidence for or against the

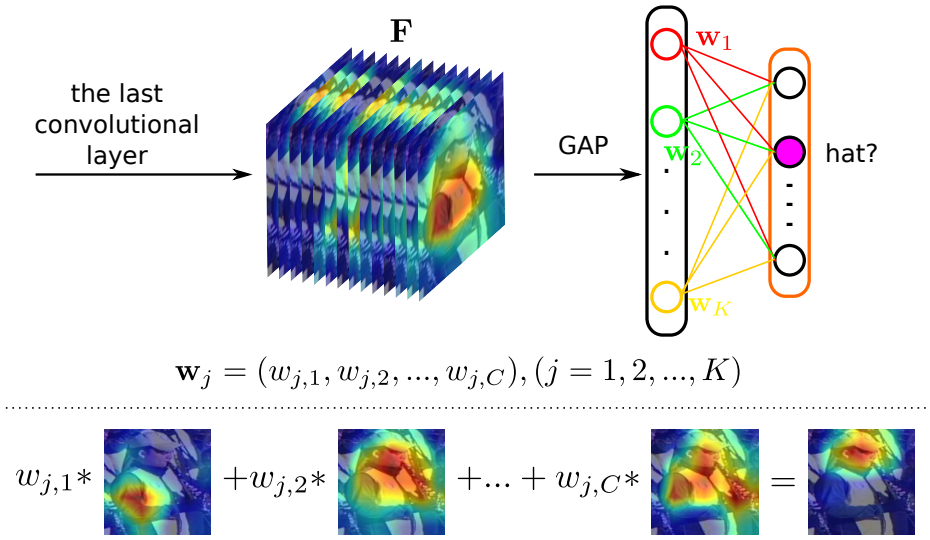


Figure 2.6 The illustration of estimating attention map based on CAM. $j$ is for the image label, i.e., human attribute, of "hat".

image label presence. Thus, the attention maps estimated for the deep network are actually byproducts of the network prediction.

# Chapter 3

# Literature Review

This chapter provides a literature review for the works related to this research, including prior works on multi-label visual recognition, visual attention mechanism and consistency-based regularization.

## 3.1 Multi-label Visual Recognition

Multi-label visual recognition, a.k.a., multi-label image classification [149, 169], is a vision-based multi-label learning task and has been widely explored, with progress on both label-separate and label-correlated methods. Label-separate methods use binary relevance strategy [5] to convert multi-label visual recognition to multiple binary image classification problem. With great success of using CNNs [70, 135, 49, 57] for single-label image classification [22], multi-label visual recognition has been improved significantly. Besides, deep convolutional ranking [40] optimizes top-$k$ ranking loss on convolutional architectures to learn a better feature representation. Hypotheses-CNN-Pooling [161] aggregates object segmentation hypotheses with max pooling to generate multi-label predictions.

Much progress has been made on label-correlated multi-label visual recognition in recent years. Many methods, such as matrix completion [7], probabilistic label enhancement [89], RGNN [173], SINN [54], Conditional Graphical Lasso [87], CNN-RNN [153], and ML-GCN [12] are proposed to model the semantic correlations between labels for multi-label visual recognition. For example, CNN-RNN [153] combines RNNs with CNNs to learn the correlations between different labels. ML-

GCN [12] adopts Graph Convolutional Networks (GCN) to embed the label correlations to the classifier learning.

Furthermore, Spatial Regularization Network [178] captures both semantic and spatial correlations between labels. Label balancing [46] is also used for improving multi-label image classification. In recent years, some sub-tasks of multi-label visual recognition, such as extreme multi-label classification [136] and partial multi-label learning [30], have attracted many research interests in computer vision field.

### 3.1.1 Human Attribute Recognition

While human attribute recognition [156] is a specific problem of multi-label visual recognition, it regards the human attributes as the image label, which depicts the person, i.e., the visual content, in the image. As the images depicting persons, e.g., pedestrians in the wild or surveillance scenarios, are usually with low quality, human attribute recognition is more challenging than ordinary multi-label image classification, and attracting more research interests recently.

Earlier methods [4, 179, 23] leverage hand-crafted features to recognize each attribute based on human appearance. As deep networks grow prosperous in the computer-vision field, Convolutional Neural Networks (CNNs) [70, 135, 140, 49, 57] have become a standard component [138, 83] and achieved great success in human attribute recognition.

Similar to the methods proposed for ordinary multi-label visual recognition, recent methods for human attribute recognition can be classified into two main categories: attribute-correlation methods [153, 154, 174, 45, 88, 141], which explore semantic attribute dependencies to facilitate the attribute recognition, and attribute-localization methods, which utilize the attribute-relevant image regions for spatially more focused attribute recognition. Some examples of attribute-correlation methods include JRL [154] utilizing LSTM [52] to explore attribute context and correlation for at-

tribute recognition, GRL [174] establishing grouping recurrent learning to leverage the attribute dependency of intra-group mutual exclusion and inter-group correlation, JLAC [141] using GCN to capture the attribute dependency and context correlation for human attribute recognition, etc. Attribute-localization methods can be further classified to two sub-categories: part-based and attention-based ones.

**Part-based localization** [77, 32, 4, 64, 170, 171, 108, 38, 90, 84, 95] usually exploits the pose estimation, body-part detections or manual annotations to specify the attribute-relevant image regions. For example, [4] decomposes the image of a person into a set of poselets with rich appearance and local pose information and human attributes are then recognized using features extracted from these poselets. In [64], a feature dictionary is built to describe possible appearance variation at each human body part, which can be used to improve part detection and human attribute recognition. In [170], following Deformable Part Models (DPM) [32], Deformable Part Desciptors (DPD) are extracted for part detection and attribute recognition. Based on CNNs, several deep part models are developed for human attribute recognition. In [171], a PANDA system leveraging CNNs trained for each poselet is developed for attribute recognition. [108] proposes an attribute grammar model to jointly represent both the object parts and their semantic attributes within a unified compositional hierarchy. [38] suggests the use of deep poselets as a part detector to localize human body parts under different poses. Deep-Context [90], another part-based method using deep learning, improves human attribute recognition by using hierarchical contexts.

In these methods, attribute recognition is usually accomplished by taking a two-step procedure. First, a body-part detector or pose estimator is applied to localize important human body parts, such as head, legs, arms, hands, neck, eyes, etc. Second, image features are extracted at each body part and then fed to pre-trained classifiers

for attribute recognition. Typically, one classifier is trained for each attribute and this classifier usually takes only the features from the corresponding body parts [4, 64, 170, 171, 108]. There is also research work that suggests the combination of features from different body parts for attribute recognition [38]. There are several issues in using part-based methods for human attribute recognition. 1) It requires prior correspondence between body parts and attributes. This may not be trivial for some attributes. 2) It requires an accurate and reliable body-part detector and/or pose estimator, both of which are well known challenging tasks in computer vision. Errors in detecting body parts can seriously hurt the performance of attribute recognition. 3) The training of body-part detector usually requires manual annotations of body parts in large-scale images and this can be highly laborious. The R*CNN method [39] does not detect body parts and consider the correspondence between body parts and attributes. Instead, in R*CNN, contextual cues in larger regions are exploited and used for facilitating human attribute recognition. The Deep-Context [90] also combines both part-based and contextual information for attribute recognition.

**Attention-based localization** [97, 178, 124, 143, 142, 165] applies spatial attention mechanisms to discover attribute-relevant image regions for improving attribute recognition. For example, HydraPlus-Net [97] hierarchically discovers the attribute relevant regions for attribute recognition. VAA [124] aggregates spatial representations by self-attention to refine the attribute recognition. ALM [143] adaptively discover discriminative regions for human attribute recognition in multi scales. DaHAR [165] uses pre-trained person segmentation as prior knowledge to exclude distractive image regions for human attribute recognition. The methods proposed in this study for human attribute recognition fall in the category of attention-based localization, for which we define and enforces two kinds of regularization, i.e., atten-

tion concentration and attention consistency, on deep network learning for human attribute recognition.

### 3.1.2 Long-tailed Multi-label Visual Recognition

Relying on label localization or label correlation, the above methods for multi-label visual recognition suffer from long-tailed training data. When the training set is long-tailed, head classes usually dominate the network training, resulting in inaccurate label localization and label correlations for tail classes, which severely hurts the recognition performance on tail classes. To address the long-tailed issue, a lot of methods are proposed for balancing the deep network training on recognizing different classes by data re-balancing. Usually, data re-balancing emphasizes tail classes more in the network learning, and it has achieved improved results on many long-tailed recognition tasks. Re-balanced sampling [10, 129, 6, 47, 176] and cost sensitive re-weighting [8, 19, 56, 158, 112, 83, 141] are the two typical kinds of data re-balancing methods. The former improves the class balance by either up-sampling the tail classes or down-sampling the head classes, while the latter improves the class balance by weighting more on tail classes in the loss functions. However, all these methods are for single-label recognition, i.e., each image only has one label. [166] extend re-balanced sampling and cost-sensitive re-weighting methods to handle long-tailed multi-label visual recognition and propose an optimized DB Focal method, which does improve the recognition performance of tail classes. However, because of label co-occurrence in multi-label recognition, emphasizing the tail classes may impair the head-class training. The re-balanced sampling may simultaneously decrease the performance of some head classes [166]. Considering that the original long-tailed training data and re-balanced training data have the distributions bias towards head classes and tail classes, respectively, this study expects to enforce the deep network to learn from a compromise distribution between these two biased distributions. The

compromising would lead to a more balanced distribution somewhere between two biased distributions and training on it would result in better multi-label visual recognition performance with long-tailed training set.

## 3.2 Visual Attention Mechanism

This section introduces the visual attention mechanism in deep networks, which imitates the human visual attention.

### 3.2.1 Human Visual Attention

According to studies on human cognitive [24, 75] and neuroscience [102], humans mainly rely on only part of an image to recognize a class in it. Intuitively, when a person wants to recognize a dog from an image, his/her attention should be attracted to the image region that depicts the dog, while the remaining regions contribute much less on the recognition process. In cognitive psychology, human visual attention is elaborated mainly in two stages [63]. In the case of image-based recognition task, the first stage is to distribute human visual attention uniformly over the entire images, called visual scene, and process the information in parallel. The second stage is to concentrate visual attention to a focused region of the image, and process the information in a serial manner. To further explain the visual attention, researchers proposed at least two models to describe the operation of visual attention, such as spotlight [27] and zoom-lens [28]. Even though debates exist between scientists on the details of human visual attention operation, there is no explicit disagreement that attended areas are much important for human recognition.

### 3.2.2 Deep Network Visual Attention

As an imitation of human attention, deep network attention tries to locate the relevant information and focus attention on it by assigning more weight on it for certain

predictions, such as the transformer [151] for natural language processing and spatial visual attention for computer vision. In this dissertation, the spatial visual attention of deep networks is mainly referred.

The visual attention of deep networks has drawn significant research interest in recent years. In general, prior works on deep visual attention can be categorized to either bottom-up attention or top-down attention. The bottom-up attention maps are learned during network forwarding. They can be used as masks to actively help the network focus on discriminative regions, such as STN [60], SENet [55], CBAM [162], and RAN [152]. The saliency [53, 1, 9, 130, 155, 13], which captures image regions that stands out from its neighbors and attract the observer's attention, can also be regarded as a kind of bottom-up visual attention. The top-down attention maps are usually inferred based on network predictions. Instead of being used for masking, top-down attention maps are more interpretative, i.e., specifying the influence of each image region to the network decisions, and therefore they are also called attribution maps [3]. In this study, the plausibility of attention maps is expected to be improved for refining the network learning, for which the top-down, interpretative attention maps are adopted.

Three categories of methods have been used for estimating the top-down attention maps [3]. Perturbation-based methods [168, 114, 34, 20, 115, 180] usually remove a portion of an input image before feeding the image through the network to infer the effect of the removed region to the network prediction. Gradient-based methods [134, 132, 139, 127] calculate the gradients in the back-propagation to quantize the contribution of input-image pixels to certain recognition. Structure-based methods [107, 175] produce attention maps by re-weighting the activation maps on the basis of certain network architectures, such as global average pooling [175]. This research uses a structure-based method, i.e., Class Activation Mapping (CAM) [175], for attention map estimation, since it is differentiable and computationally efficient.

## 3.3 Consistency-based Regularization

Consistency is an important property in computer vision field for regularizing the deep network learning. Different kinds of network consistency have been considered for improving network training in different tasks. *Perturbation-based consistency* requires a trained network to produce same prediction after applying a small perturbation to the input image [131, 123, 101, 163, 159] and it has been widely used for data augmentation [131]. *Model-based consistency* [72, 167, 172, 109, 105] is usually formulated and applied between networks. It enforces the two different networks to produce the same results when the same image is taken as the input. Examples include Π-model [72] and Mean Teacher [144] used for semi-supervised learning, deep mutual learning [172] and co-regularization [105] for training two networks collaboratively, and co-teaching [43] for handling noisy labels. However, by taking the input from the same distribution, two branches trained in [172, 105] may collapse to each other if their network parameters are not carefully initialized with substantial difference. In [109], an adversarial scheme is introduced to address this issue.

To address the label locality in human attribute recognition, two kinds of attention consistency are adopted in this research: the perturbation-based equivariance under spatial transforms and the model-based invariance between different networks. For the long-tailed issue of multi-label visual recognition, the model-based consistency is leveraged to make the proposed network compromise between two biased distributions.

### 3.3.1 Deep Equivariance

Equivariance is studied as an important mathematical property [81] of spatial representations. It indicates that certain spatial representations of images should follow the same transform if the image is spatially transformed. Some image representations, such as HOG [21], have been proved to be adhering to this property [81]. Previous

works also attempt to construct equivariant representations [51, 67, 126], including the deep convolutional representations with certain equivariance [81]. Most of existing works on applying equivariance to deep network learning are focused on the features at certain convolutional layers [82, 25, 17, 164, 146, 145, 101, 110, 163]. Differently, this study proposes the attention equivariance of deep networks at a specific semantic stage of the networks – the estimated attention maps reflecting the local spatiality of the considered human attribute.

### 3.3.2 Deep Invariance Cross Networks

Enforcing consistency between different networks can be regarded as a kind of collaborative learning. Following the principle that "Two heads are better than one", the training of a network can be regularized by transferring information learned by another network, as verified in the research on knowledge distillation [50, 167, 144]. Different kinds of collaborative learning, e.g., deep mutual learning [172], co-regularization [105] and co-training [109], have been studied to transfer information between two different networks, leading to regularized training of both networks. Most of them try to minimize the final prediction difference between networks for the same input. Each network provides smoothed ground truth for the training of other network, leading to better network performance [103]. Co-teaching [43, 100] also learns two networks simultaneously, with each network helping the other by excluding samples with noisy labels, i.e., it uses consistency of predictions for sample selection, instead of prediction supervision. Since the final prediction is aggregated for the whole image, all these methods lack consideration of local spatiality in defining the cross-network consistency. Thus, the attention consistency of invariance between different networks is proposed to explicitly consider local spatiality that is important for human attribute recognition during collaborative learning. Besides, the collaborative

learning is also adopted to compromise the learning between two biased distributions for addressing the long-tailed issue in multi-label visual recognition.

# Chapter 4

# Visual Attention Concentration for Focused Attribute Recognition

To address the locality of labels (human attributes), it is important to have the deep network focus attention on attribute regions for human attribute recognition. This goal is achieved by enforcing the concentration of deep network visual attention in this study. Compared with part-based methods, the proposed attention concentration does not require 1) manually annotated human body parts or pre-trained human body part detector, which is very challenging; and 2) prior correspondence between attributes and human body parts, which can not be well defined for certain attributes.

The proposed attention concentration uses Class Activation Mapping (CAM) [175] to estimate attention maps for recognizing each human attribute. Then, a new exponential loss function is proposed to measure the concentration of each attention map. The deep network is trained in terms of both the original image classification loss and the proposed exponential loss. To verify the effectiveness of the proposed method, experiments are conducted on Berkeley Attributes of Human People Dataset [4] and WIDER Attribute Dataset [90]. The following bullet points convey the core findings of this study.

- The concentration of attention maps reflects the generalization ability of the deep network for attribute recognition.
- Deep networks with CAM can be improved for refining the visual attention by including a new component.

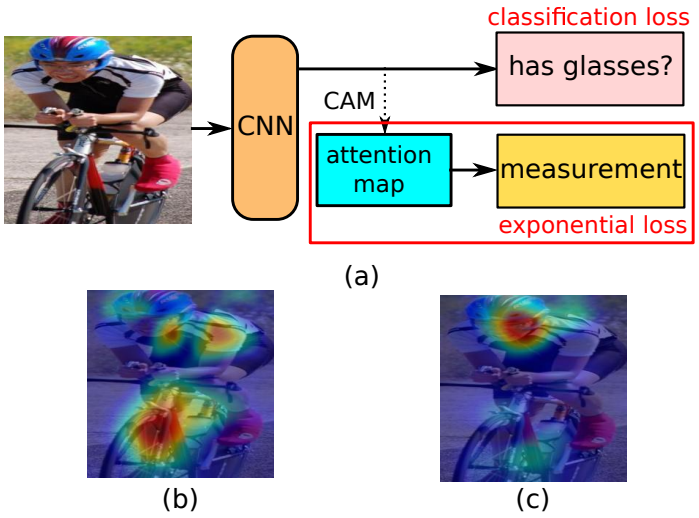


Figure 4.1 Motivation illustration of the proposed method. (a) The deep network and the added component for refining attention maps. (b) and (c) Attention map before and after the refining for recognizing the same attribute, respectively.

- Exponential loss function is a reasonable choice for measuring the appropriateness of visual attention of deep networks.

## 4.1 Overview

Even though part-based methods for human attribute recognition are limited by not well defined correspondence between body parts and human attributes, inaccurate body-part detector and/or pose estimator, intense labor for manually annotating body parts in large-scale image dataset, they demonstrate that certain image regions play the most import role in recognizing a human attribute. This verifies the assumption of focusing on attribute-relevant regions benefits the deep networks for human attribute recognition. To achieve this focused human attribute recognition, instead of using body-part detectors, this study proposes to automatically identify attribute-relevant regions for each attribute and enhances the visual attention concentration of deep networks for human attribute recognition.

Specifically, as shown in Fig. 4.1(a), a CNN is trained to recognize a human attribute from an input image. Meanwhile, the attention map for recognizing this attribute is estimated during the network inference. Originally, the attention map could highlight semantically irrelevant image regions for recognizing this attribute, i.e., clothes and the bicycle tie areas for recognizing the attribute "glass", as shown in Fig. 4.1(b). To address this issue, an additional component is introduced to measure the concentration of the attention map for refinement. As shown in Fig. 4.1(c), after enhancing the attention concentration, the attention map is highly focused on the attribute-relevant regions, i.e., face regions.

## 4.2 Methodology

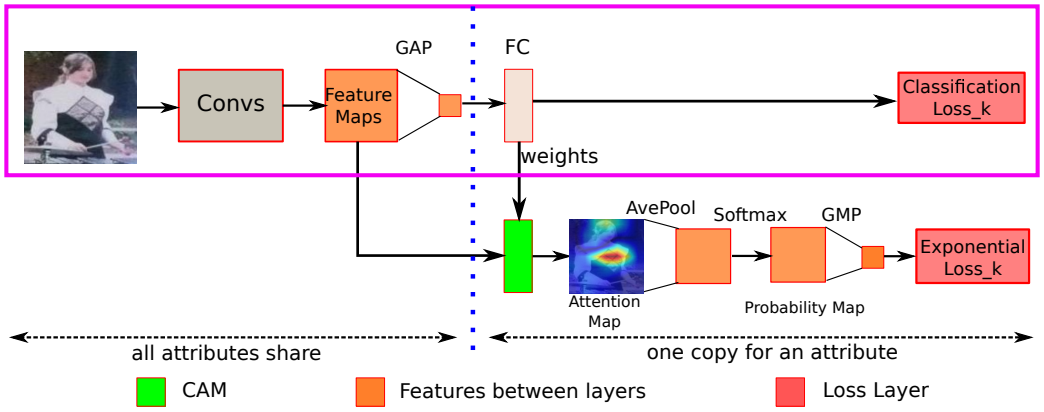


Figure 4.2 An illustration for the framework of the proposed method.

As shown in Fig. 4.2, the proposed method adds an additional component for the deep network learning when recognizing each human attribute. Modules in the pink box construct a deep network, e.g., AlexNet-CAM or VGG16-CAM. An input image first goes though the convolutional module of the networks, denoted by "Convs" in the figure. The predicted attribute presence score is fed to the cross-entropy-based classification loss in Eq. (2.4). Meanwhile, the attention map for recognizing each attribute is estimated by CAM, discussed in Section 2.4.

In the ideal case, the attention map highlights the most relevant regions for the considered attribute. However, in practice, overly small size and low image quality of the actual relevant regions and over-fitting training (mainly due to insufficient training data) may lead to incorrect attention maps. Examples are shown in Fig. 4.3, where a set of image and their attention maps are arranged side by side, with the considered attribute labeled at the bottom left corner of each image. The results show that the computed attention map may be incorrect by highlighting irrelevant regions. For example, the highlighted area is not focused on face region when recognizing the attribute of glasses in the image at the center of Fig. 4.3.


Figure 4.3 Sample images and the corresponding attention maps, which may not highlight the correct regions for the considered attribute.

### 4.2.1 Attention Concentration Enhancement

This study proposes to enhance the concentration of attention maps for human attribute recognition. As shown in the lower part of the framework in Fig. 4.2, a new component consisting of several layers is added to the deep network. First, this added component includes a CAM module to compute the attention map based on the class activation mapping [175] as described in Section 2.4. Then, an average pooling layer is included to capture the importance of all the potential relevant regions for recognizing the considered human attribute. A 2D softmax function is used to convert the pooled attention maps to a probability map: Let $z(m, n)$ be a value of the location $(m, n)$ in a pooled attention map and the corresponding value in the probability map is computed as

$$Softmax_{2D}(z_{m,n}) = \frac{e^{z_{m,n}}}{\sum_{p=1}^{H}\sum_{q=1}^{H} e^{z_{p,q}}}, \tag{4.1}$$

where the size of the pooled attention map is $H \times H$. Finally, a Global Maximum Pooling (GMP) layer is included to extract the maximum probability, which reflects the credibility of the identified relevant region. Based on this maximum probability, a loss function is defined to reflect the concentration of current attention maps. Two key issues need to be addressed in this component: 1) the definition of the loss function, which measures the concentration of an attention map, based on the maximum probability, and 2) the tuning of the deep network to increase the maximum probability and minimize the loss function.

### 4.2.2 Exponential Loss

Due to the softmax function, the summation over the probability map is one. Thus, increasing the maximum probability will automatically suppress the regions with smaller probability. This will make the attention region (highlighted region) in the attention map more concentrated, which increases the chance of capturing the region relevant to the considered attribute. In this study, the exponential loss function is

developed independent of any supervised information on the attention map, based on the maximum probability on the probability map. Let $p_{x,j}^M$ be the maximum probability for image $\mathbf{x}$ and attribute $j$. The loss function for $j$th-attribute is defined as

$$\mathcal{L}_{con} = \frac{1}{K} \sum_{j=1}^{K} e^{\alpha(p_{x,j}^M + \beta\mu)}, \tag{4.2}$$

where $\alpha$ and $\beta$ are adjustable parameters of the loss function, $\mu = 1/H^2$ is the mean value of the probability map, with $H \times H$ being the size of the attention maps and the probability maps, and $K$ is the number of human attributes.

Given that the attention map size $H \times H$ is fixed if the input image size is fixed, $\mu$ is also a constant. Since the loss function $\mathcal{L}_{con}$ is negatively related to the maximum probability $p_{x,j}^M$, $\alpha$ is a negative parameter. Furthermore, $\beta$ also takes a negative value, making $|\beta\mu|$ a threshold: If the probability $p_{x,j}^M$ is less than this threshold, the loss value is large, indicating that the attention map is not concentrated.

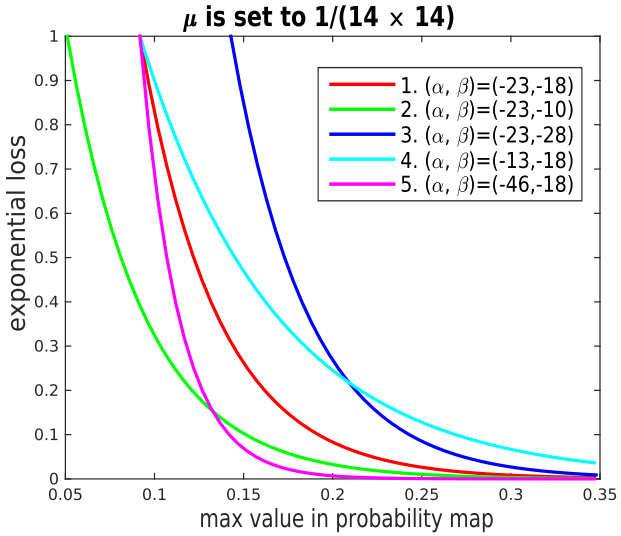


Figure 4.4 Curves of the proposed exponential loss function – the loss decreases with the increase of the maximum probability.

Figure 4.4 shows the curves of this loss function over an attention map. Curves 1, 2, and 3 share a same $\alpha$ but different $\beta$. Curves 1, 4, and 5 share a same $\beta$, but

different $\alpha$. It can be noticed that $\alpha$ controls the descent rate of the loss value, while $\beta$ adjusts the impact of the maximum probability by a threshold – the maximum probability $p_{x,j}^M$ takes value in the range $[\mu, 1]$. Based on the empirical observation, given $\mu = 1/(14 \times 14)$, the attention region is highly concentrated when $p_{x,j}^M$ is above 0.2. Therefore, it is desired that the loss $\mathcal{L}_{con}$ becomes very small when $p_{x,j}^M \geq 0.2$. This is used to guide the selection of $\alpha$ and $\beta$ in the experiments.

### 4.2.3 Network Tuning

As described above, on the proposed network illustrated in Fig. 4.2, two loss functions exist – the original *classification loss function* aiming at reducing the attribute recognition error rate and the added *exponential loss function* aiming at enhancing the attention maps. The proposed network is trained in two steps: 1) *Pre-training*. In this step, without considering the added component and the exponential loss function, i.e. the modules included in the pink box in Fig. 4.2, the original deep network is trained by minimizing classification loss function. 2) *Fine-tuning*. In this step, this study fine-tunes the network parameters by minimizing both loss functions. Note that all the layers in the added component outside the pink box in Fig. 4.2 do not have free parameters to tune. The exponential loss is back propagated through the added component to fine-tune the parameters of the convolutional layers in the CAM network. In the meantime, the classification loss function is also back propagated through the fully connected layer and then convolutional layers (in the pink box in Fig. 4.2) to fine-tune their parameters. This way, the parameters of convolutional layers are actually fine-tuned by minimizing a loss function that combines the exponential loss function and the classification loss function through back propagation.

## 4.3 Experiment

The proposed network is built on the Caffe [62] platform, by customizing a *weighted average layer* and an *exponential loss layer*. The customization of the weighted average layer is used to achieve CAM and embed the attention map into the network training. The Berkeley Attributes of Human People Dataset [4] and the WIDER Attribute Dataset [90] are used to train the proposed networks and evaluate the proposed method, individually.

The Berkeley Attributes of Human People Dataset contains 8,035 images, each of which is centered at a full body of a person. Nine human attributes are referred in this dataset, including "is male", "long hair", "glasses", "has hat", "has t-shirt", "long sleeves", "has shorts", "has jeans" and "long pants". This dataset is divided into two subsets: the training subset with 4,013 images and the testing subset with 4,022 images.

The WIDER Attribute Dataset contains 13,789 images with 57,524 annotated persons, each labeled with 14 human attributes, including "male", "long hair", "sunglasses", "hat", "t-shirt", "long sleeves", "formal", "shorts", "jeans", "long pants", "skirt", "face mask", "logo" and "stripe". It is divided into 5,509 training, 1,362 validation and 6,918 testing images (13,789 in total). The training and validation subsets is used for training and the testing subset for testing.

**Pre-Training:** In the experiment, networks are constructed from classical AlexNet and VGG16, with replacing the final FC stack with GAP-FC structure and denoted as AlexNet-CAM and VGG16-CAM, respectively, as discussed in Section 2.3. This study pre-trains the networks using the following three steps. 1) Taking the existing AlexNet and VGG16 models pre-trained on ImageNet/ILSVRC [121]. 2) Further training AlexNet and VGG16 using training samples in Berkeley Attributes of Human People Dataset or WIDER Attribute Dataset, separately. 3) Training AlexNet-CAM

and VGG16-CAM (with the pink box in Fig. 4.2) using training samples in Berkeley Attributes of Human People Dataset or WIDER Attribute Dataset, separately. In Step 2), as shown in Fig. 2.5, all attributes share the same convolutional moduls from either AlexNet or VGG16 network, but use distinct linear operations, i.e., $FC_1$, $FC_2$, ..., $FC_K$. In the training, each classifier in such an FC layer is updated independently from other classifiers, by simply using the standard back-propagation algorithm. In Step 3), base learning rate 0.0001 is used for both networks and the iterative training ends when the classification loss function decreases to an order of magnitude of $10^{-4}$.

**Attention Concentration Fine-tuning:** After the pre-training, the classification loss is usually low, e.g., $10^{-4}$. At this stage, the exponential loss is much higher, e.g., $10^{-1}$. It is necessary to adjust the control parameters $\alpha$ and $\beta$ in Eq. (4.2) such that the combined loss is not dominated by any one of them. Empirically, this study set $\alpha = -23$ and $\beta = -18$ for all the experiments. In fine-tuning with the added component, gradients resulting from the back propagation of the two loss functions are simply added to update the parameters of the convolutional layers. The trained models are denoted as AlexNet-CAM-AC and VGG16-CAM-AC when using AlexNet-CAM and VGG16-CAM, respectively. After average pooling, each pooled attention map is a square matrix, while the Softmax layer requires an input of vector. To address this issue, the square matrix are flattened to a vector by concatenating all the rows. After the probabilities are computed, the resulting vector is converted back to a square matrix to form a probability map.

In addition, for each attribute from each image, there are three possible status: "positive", "negative" and "non-specified". "Positive" indicates the presence of the attribute in the considered image, while "negative" indicates the non-presence. "Non-specified" indicates that it is unknown whether the attribute is present or not in the image. In the experiments, if an attribute is "non-specified" for an image, this image

will not be included to train the network for this particular attribute. Besides, the added component for enhancing attention concentration is only used in training the networks. In the testing stage, only the network shown in the pink box in Fig. 4.2 is used to determine whether an attribute is present in a testing image or not.

### 4.3.1 Quantitative Results

This study first tests the effectiveness of the proposed method on the Berkeley Attributes of Human People Dataset. Table 4.1 shows the mean Average Precision (mAP) of the proposed methods, AlexNet-CAM-AC and VGG16-CAM-AC, by adapting CAM with the added attention concentration enforcement. The mAPs of AlexNet and VGG16 (mAP of VGG16 is cited from [38]), trained using Steps 1) and 2) in the above-mentioned network pre-training and mAPs of the AlexNet-CAM and VGG16-CAM, trained using all three steps in the above-mentioned pre-training are also reported. Notice that VGG16 performs much better than AlexNet, especially on the attributes of "glasses" and "hat". Using either AlexNet or VGG16, the constructed CAMs always lead to improved mAPs and the proposed added component for enforcing attention concentration can further improve the mAP performance. Figure 4.5 compares the original attention maps and the ones with concentration enforced by the proposed method for several sample images. All of them are based on VGG16 networks. It can be found that using the proposed method, the obtained attention maps are more concentrated on the desired relevant region when recognizing an attribute.

An experiment is also conducted to justify the steps in the added attention concentration component, which consists of 1) CAM module, 2) average pooling, 3) softmax, 4) global max pooling and 5) exponential loss, as shown in Fig. 4.2. Among them, step 1) calculates the attention maps from the network. Step 3) converts the attention map to the probability maps to emphasize relevant regions while automatically suppressing the irrelevant regions. Step 4) and 5) specify the loss, which is

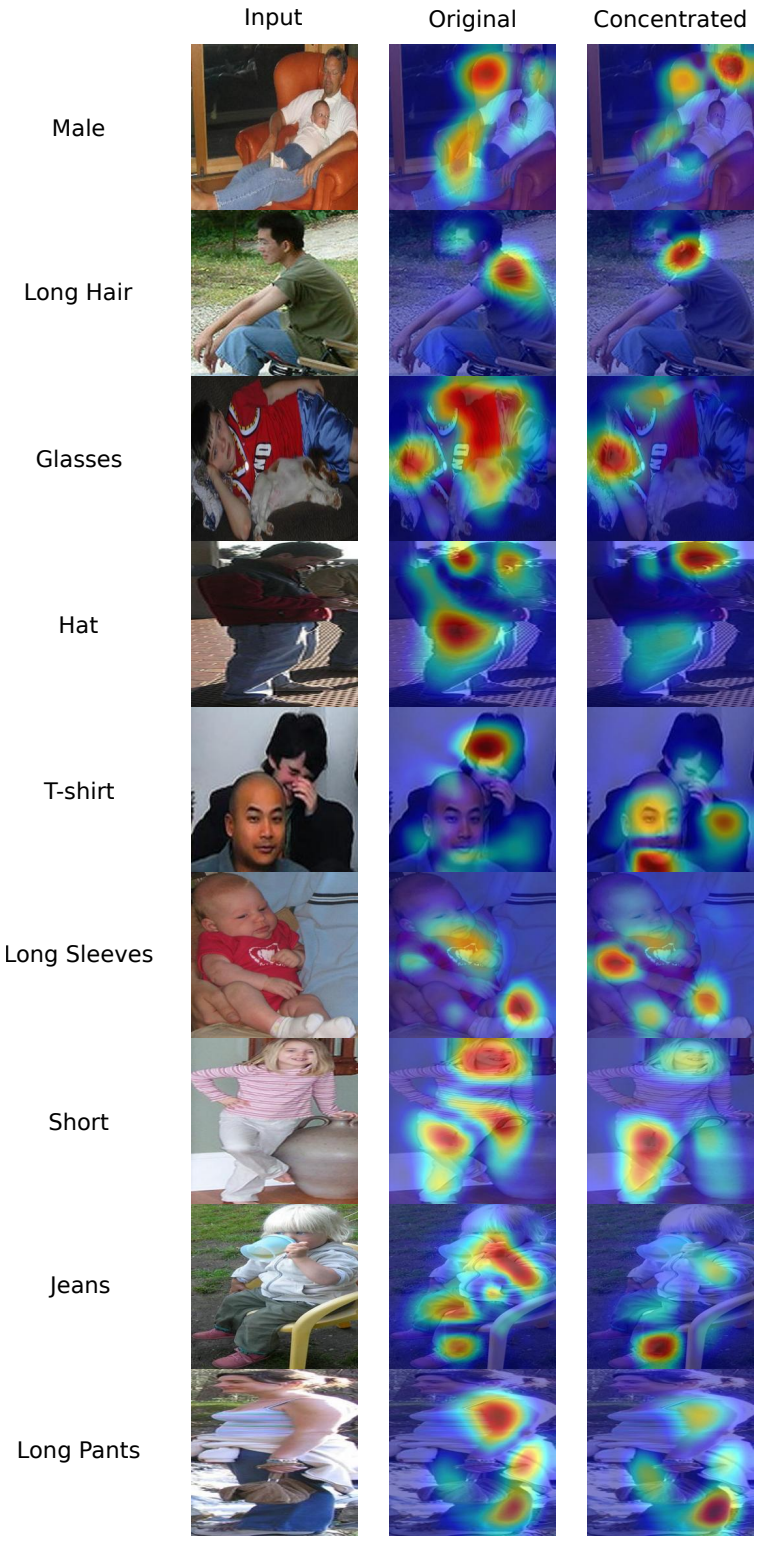


Figure 4.5 Sample results of attention map concentration. Left column: input images and the considered human attribute. Middle column: the original attention maps from VGG16-CAM. Right column: the concentrated attention maps from VGG16-CAM-AC.

Table 4.1 Attribute recognition performance of AlexNet, VGG16, AlexNet-CAM, VGG16-CAM and the proposed methods on Berkeley Attributes of Human People Dataset.

| AP(%) | *male* | *long hair* | *glasses* | *hat* | *t-shirt* | *long sleeves* | *shorts* | *jeans* | *long pants* | mAP |
|---|---|---|---|---|---|---|---|---|---|---|
| AlexNet | 84.9 | 76.0 | 46.1 | 76.1 | 60.3 | 86.7 | 86.9 | 87.5 | 97.1 | 78.0 |
| AlexNet-CAM | 88.6 | 82.4 | 55.6 | 83.1 | 65.7 | 89.0 | 88.4 | **90.0** | 98.0 | 82.3 |
| AlexNet-CAM-AC | **88.7** | **83.0** | 56.9 | **83.8** | **67.7** | **89.2** | **89.7** | 89.5 | **98.3** | **83.0** |
| AlexNet-CAM-AC w/o AvePool | 88.3 | 82.6 | **57.1** | 83.5 | 67.6 | 89.1 | 89.5 | 89.8 | 98.0 | 82.8 |
| VGG16 | 93.4 | 88.7 | 72.5 | 91.9 | 72.1 | **94.1** | 92.3 | 91.9 | **98.8** | 88.4 |
| VGG16-CAM | 93.5 | 90.7 | 76.7 | **93.8** | 75.3 | 92.7 | 92.1 | 92.5 | 98.3 | 89.5 |
| VGG16-CAM-AC | **94.1** | **90.8** | **79.6** | 93.3 | **77.2** | 93.2 | 92.1 | **92.8** | 98.6 | **90.2** |
| VGG16-CAM-AC w/o AvePool | 93.6 | 90.7 | 77.2 | 93.2 | 76.6 | 93.4 | **92.8** | 92.5 | **98.8** | 89.9 |

required for any learning framework. These four steps cannot be removed. When step 2) is removed in the proposed method, the performance is shown in Table 4.1 as 'AlexNet-CAM-AC w/o AvePool' and 'VGG16-CAM-AC w/o AvePool'. The results show that the inclusion of average pooling in the attention concentration component does improve the attribute recognition performance.

To further justify the effectiveness of the proposed method, this study also compares the performance of the proposed method against part-based methods for human attribute recognition. Specifically, Poselet [4], Deformable Part Descriptors (DPD) [170], Joo et al. [64], PANDA (Pose Aligned Networks for Deep Attribute) [171], Park et al. [108], Gkioxari et al. [38], Gkioxari et al. [39] and Deep-Context [90] are chosen for comparison. Among these eight comparison methods, the first six methods are part based and the seventh one, i.e., Gkioxari et al. [39], is a contextual cues based, while the eighth one, i.e., Deep-Context [90], is both part and context based approach for human attribute recognition. Table 4.2 summarizes the mAPs of these methods and the proposed method (VGG16-CAM-AC) on the testing data

of the Berkeley Attributes of Human People Dataset. For all the comparison methods, their mAP performance are directly copied from their respective papers. The comparison results show that, VGG16-CAM-AC achieves second best mAP of 90.2%, while the best mAP is 92.2% from Deep-Context [90].

Table 4.2 mAP performance of the proposed method and eight comparison methods on Berkeley Attribute of Human People Dataset.

| AP(%) | *male* | *long hair* | *glasses* | *hat* | *t-shirt* | *long sleeves* | *shorts* | *jeans* | *long pants* | mAP |
|---|---|---|---|---|---|---|---|---|---|---|
| Poselet [4] | 82.4 | 72.5 | 55.6 | 60.1 | 51.2 | 74.2 | 45.5 | 54.7 | 90.3 | 65.2 |
| DPD [170] | 83.7 | 70.0 | 38.1 | 73.4 | 49.8 | 78.1 | 64.1 | 78.1 | 93.5 | 69.9 |
| Joo et al. [64] | 88.0 | 80.1 | 56.0 | 75.4 | 53.5 | 75.2 | 47.6 | 69.3 | 91.1 | 70.7 |
| PANDA [171] | 91.7 | 82.7 | 70.0 | 74.2 | 49.8 | 86.0 | 79.1 | 81.0 | 96.4 | 79.0 |
| Park et al. [108] | 92.1 | 85.2 | 69.4 | 76.2 | 69.1 | 84.4 | 68.2 | 82.4 | 94.9 | 80.2 |
| Gkioxari et al. [38] | 92.9 | 90.1 | 77.7 | 93.6 | 72.6 | 93.2 | **93.9** | 92.1 | 98.8 | 89.5 |
| Gkioxari et al. [39] | 92.8 | 88.9 | 82.4 | 92.2 | 74.8 | 91.2 | 92.9 | 89.4 | 97.9 | 89.2 |
| Deep-Context [90] | **95.0** | **92.4** | **89.3** | **95.8** | **79.1** | **94.3** | 93.7 | 91.0 | **99.2** | **92.2** |
| VGG16-CAM-AC | 94.1 | 90.8 | 79.6 | 93.3 | 77.2 | 93.2 | 92.1 | **92.8** | 98.6 | 90.2 |

The proposed method is also tested on the WIDER Attribute Dataset introduced by Deep-Context [90]. The results are reported in Tables 4.3 and 4.4. The experiments results show that the proposed methods achieve better performance than Deep-Context [90] on the WIDER Attribute Dataset. Besides, different from the proposed method, Deep-Context [90] considers the contextual information besides the relevant regions considered in the proposed method. It is reasonable to believe that the proposed method can be enhanced by further considering contextual information as in Deep-Context.

### 4.3.2 Qualitative Analysis

Figure 4.6 illustrates enforcing attention concentration changes an attention map. It verifies the effectiveness of using both loss functions, i.e., the classification loss func-

Table 4.3 Comparing mAP performance on the test set of WIDER Attribute Dataset.

| Methods | mAP(%) |
|---|---|
| R-CNN [36] | 80.0 |
| R*CNN [39] | 80.5 |
| Deep-Context [90] | 81.3 |
| VGG16 | 81.7 |
| VGG16-CAM | 82.5 |
| VGG16-CAM-AC | **82.9** |

Table 4.4 AP performance for each attribute on the test set of WIDER Attribute Dataset

| AP(%) | VGG16 | VGG16-CAM | VGG16-CAM-AC |
|---|---|---|---|
| male | 94.9 | 95.0 | **95.3** |
| long hair | 83.8 | 84.7 | **85.2** |
| sunglasses | 70.1 | 69.7 | **71.3** |
| hat | 92.5 | 93.5 | **93.6** |
| tshirt | **77.8** | 77.3 | 77.7 |
| long sleeves | 95.0 | 95.3 | **95.5** |
| formal | 78.5 | 80.6 | **80.7** |
| short | **89.3** | 88.3 | 88.9 |
| jeans | 72.5 | 74.1 | **74.9** |
| long pants | 96.2 | 96.2 | **96.3** |
| skirt | 79.2 | 80.2 | **80.7** |
| face mask | 70.1 | 72.3 | **72.6** |
| logo | 87.6 | **88.0** | 87.5 |
| stripe | 56.4 | **60.0** | **60.0** |
| mAP | 81.7 | 82.5 | **82.9** |

tion and the exponential loss function, in the proposed method. Figure 4.6(a) shows an image where we want to recognize the attribute of "has glasses". Figure 4.6(b) shows the attention map from VGG16-CAM and this attention map contains multiple highlighted regions, among which there is only one, indicated by a red circle, is the real relevant region for the considered human attribute. This is actually the result from minimizing the classification loss function. If the network only uses the exponential loss function, the fine-tuning is totally unsupervised – the resulting at-

tention map will be highly concentrated on one small region, but this region may not be relevant to the considered human attribute. By using both loss function, the concentrated attention map, as shown in Fig. 4.6(c), can get not only more concentrated but also more attribute relevant. The process can be explained as emphasizing the attention on the relevant regions while suppressing the attention on the irrelevant regions.

Additionally, in some examples in Fig. 4.5, the most highlighted area completely "moves" from one region to another region by enforcing attention concentration. This is achieved by changing the values of of the attention map. As shown in Fig. 4.6(b), both regions **A** and **B** are highlighted, although A has the highest map values, before the attention concentration. With the proposed attention concentration, the values in region B increase to make region B the most highlighted region, while the values in region A are decreased by the suppression. This leads to a visual effect of the moving the highlighted area, Fig. 4.6(c).

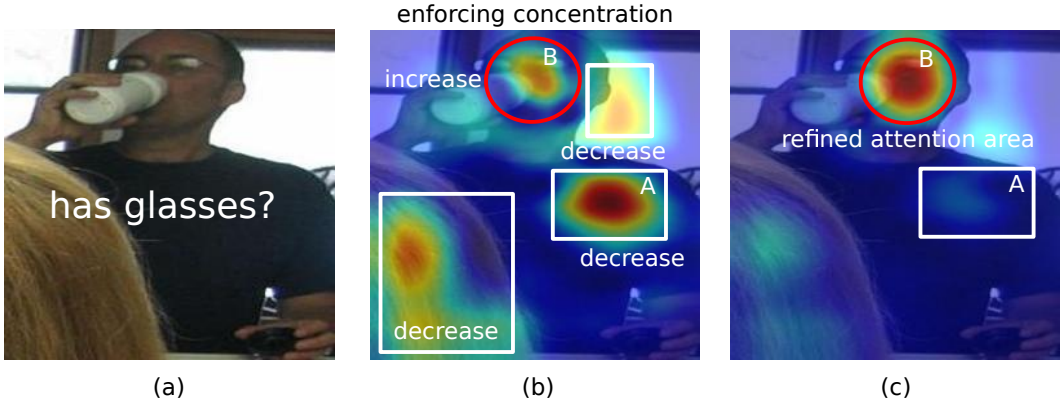


Figure 4.6 An example for illustrating the effectiveness of the two loss functions in the proposed method. (a) An image for recognizing the attribute of "has glasses"; (b) Attention map extracted by VGG16-CAM; (c) Concentrated attention map extracted from VGG16-CAM-AC.

## 4.4 Chapter Summary

To summarize, this study proposed to integrate class activation mapping for human attribute recognition, without requiring prior correspondence between the human body parts and the attributes. Based on the CAM, it further introduced a new component to extract and enhance the attention maps for each training image. A new exponential loss function was defined to measure the concentration of the attention maps. Considering this new loss function and the original classification loss function, the proposed method can highlight the attribute-relevant regions with higher concentration in the network training. This study also compared the performance of the proposed method with previous part-based attribute recognition methods on the Berkeley Attributes of Human People Dataset and WIDER Attribute Dataset. The results verified that enforcing visual attention concentration in deep network learning outperforms the part-based methods for human attribute recognition.

# Chapter 5

# Visual Attention Consistency for Consistent Attribute-Region Relevance

Since attention maps reflect the attribute-region relevance in the view of deep networks, the plausibility of attention maps indicate the attribute locality addressed by deep network learning. In this research, another method to address the attribute locality for human attribute recognition is to enforce the visual attention consistency when deep networks are learned for recognizing attributes. Consistency is an important property for robust vision systems. For example, human visual perception shows good consistency for visual recognition from images, i.e., when an image goes through certain image transforms, such as flipping, scaling and rotation, the human perception for attribute "sunglasses" appearing in the image remains consistent. This consistency has motivated the *data augmentation* strategy [70], which has been widely used in training deep networks – for each original image with ground-truth annotations, a new training image could be constructed by transforming this image and assigning the same ground-truth annotations. Data augmentation regularizes the deep network models by reducing the over-fitting problem for recognition tasks with perceptual consistency under spatial transforms.

To be specific, this study explores and enforces two kinds of attention consistency in network learning for human attribute recognition. One kind of consistency enforces the equivariance of the attention map when the input image undergoes certain spatial transforms, such as scaling, rotation and flipping. The other kind of the consistency is

enforced between the attention maps derived from two different networks when both of them are trained for recognizing the same attribute from the same image. These two kinds of consistency are formulated as new loss functions and combined with the traditional classification loss for attribute recognition learning. The proposed methods are evaluated on three representative datasets for human attribute recognition: WIDER Attribute [90], PA-100K [97], and RAP [85]. The experimental results verify the effectiveness of each of the two kinds of proposed attention consistency as well as the combination of them. The proposed methods achieve new state-of-the-art performances on these datasets.

## 5.1 Overview

As mentioned above, straightforward supervision on attention maps is not viable, since the manual annotation for ground-truth attention maps specifying attribute relevant regions is infeasible. To improve the plausibility of attention maps without using the ground-truth attention maps, this dissertation studies the consistency of the attention maps when recognizing an attribute in an image. Specifically, two kinds of attention consistency are considered: *equivariance under spatial transforms* and *invariance between different networks*. For the former, from a well trained network, the attention map of the same attribute in the same image shall be equivariant to certain spatial image transforms, i.e., if the input image undergoes a rotation, flipping or scaling transform, the attention maps derived from the network shall show the same transform to capture the consistency of attribute-relevant regions. For the latter, when two different networks are well trained for human attribute recognition, they shall produce identical attention maps when recognizing the same attribute in the same image, since the underlying attribute-relevant regions, even if difficult to manually annotate sometimes, is a visual perception concept independent of the adopted network. However, neither of these two kinds of consistency is well preserved

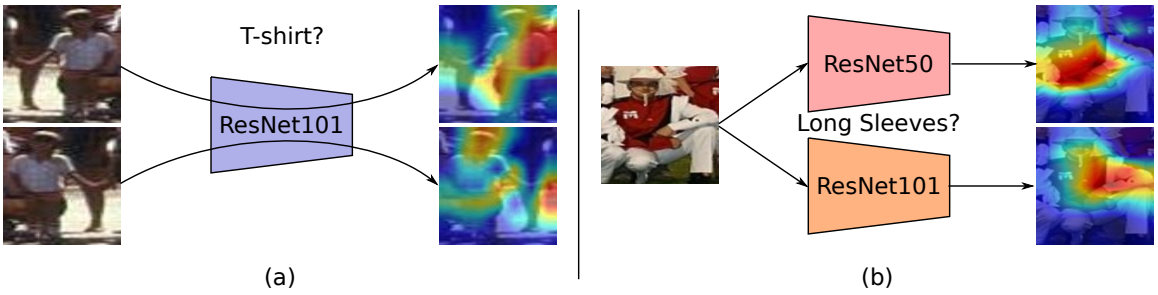


Figure 5.1 An illustration of visual attention inconsistency in the current networks for human attribute recognition. (a) In recognizing the attribute "T-shirt" using a ResNet101 [49], the flipping of the input image does not lead to the flipping of the attention map. (b) In recognizing the attribute "Long Sleeves" in an image, two networks, ResNet50 and ResNet101, produce different attention maps.

in the current deep neural networks learned for human attribute recognition, as shown in Fig. 5.1. Thus, this study develops a new approach to enforce these two kinds of attention consistency in the network training for better human attribute recognition.

To achieve these two kinds of attention consistency, a two-branch framework is proposed, where both branches are deep networks learned to recognize the same set of human attributes by minimizing cross-entropy-based image classification loss. Meanwhile, Class Activation Mapping (CAM) [175] is used to estimate attention maps for each branch. For the attention consistency of equivariance under spatial transforms, the same network with shared parameters is trained for the two branches, while the input image of one branch is spatially transformed as the input of the other branch. Then, a new attention consistency loss is defined to measure the difference between the attention maps of two branches after applying the inverse spatial transform to the attention maps of the transformed image. For the attention consistency of invariance between two networks, different networks are learned for two branches, which take the same image as input. The new attention consistency loss is also used to measure the difference of the two attention maps for recognizing the same attribute. Finally, the combination of two kinds of consistency is considered

by using two different networks for the two branches, and spatially transforming the input of one branch as the input of the other branch. In each case, the defined new consistency loss is added to the original classification losses for training the respective networks for human attribute recognition.

## 5.2 Methodology

### 5.2.1 Overview

As defined in Section 2.3, the human attribute recognition tells the presence of each human attribute from an input image $\mathbf{x} \in \mathbb{X}$ of a person. The ground-truth attribute annotations for the image are denoted as $\mathbf{y} \in \mathbb{Y}$, with $\mathbf{y} = \{y_1, y_2, ..., y_K\}$ where $y_j = 1$ if attribute $j$ is present in the image and $y_j = 0$ otherwise. $K$ is the number of considered attributes. $\mathbb{X}$ is the set of $N$ training images and $\mathbb{Y}$ is their corresponding set of ground-truth annotations.

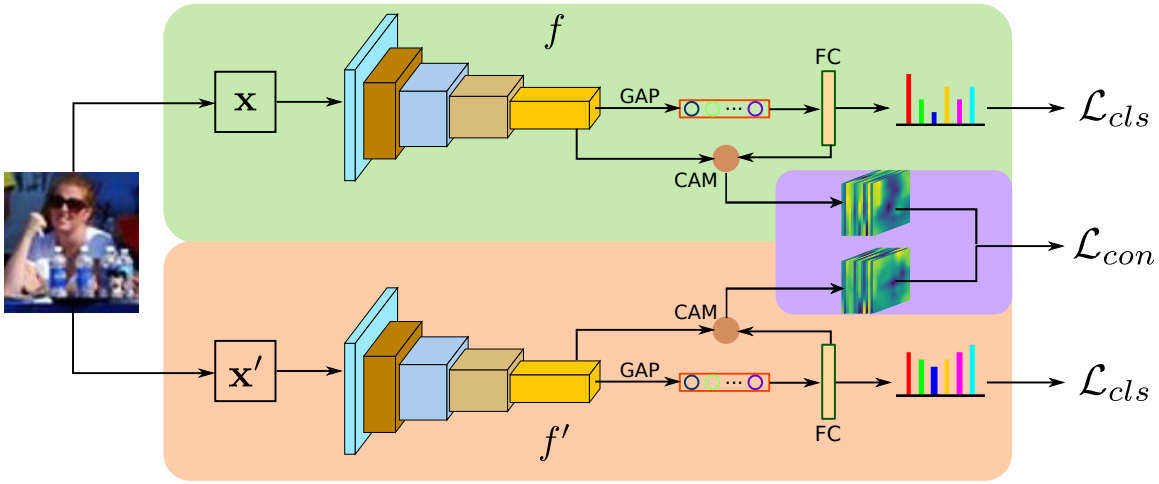


Figure 5.2 An illustration of the proposed two-branch framework.

Generally, as shown in Fig. 5.2, the proposed framework consists of two branches. Both of them are deep networks starting with convolutional layers and ending with GAP-FC (fully connected layer after global average pooling) structure, e.g., ResNet,

DenseNet [57]. The traditional binary cross entropy loss is used as the classification loss $\mathcal{L}_{cls}$ to learn each branch for recognizing the same set of human attributes. Based on the structure-based attention mechanism, CAM [175] is adopted to estimate attribute-specific attention maps for each branch. To enforce the attention consistency between two branches, a new attention consistency loss $\mathcal{L}_{con}$ is introduced based on pixel-level distance between attention maps for recognizing the same attribute in an image.

Let $\mathbf{x}$ and $\mathbf{x}'$ be the inputs, $f$ and $f'$ be the networks of the two branches, respectively. By defining them in different ways, this two-branch framework can be used to enforce the proposed two different kinds of consistency, respectively:

(a) To enforce the attention consistency of equivariance under spatial transforms, $\mathbf{x}$ and $\mathbf{x}'$ are set as the original and transformed images, respectively, i.e., $\mathbf{x}' = T(\mathbf{x})$, where $T$ is a spatial transform, such as flipping, scaling and rotation. Besides, the networks in two branches are identical, sharing the architecture and parameters, i.e., $f' = f$. The attention map estimated on $\mathbf{x}'$ goes through the inverse transform $T^{-1}$ before being compared to the attention map of $\mathbf{x}$ for the calculation of attention consistency loss $\mathcal{L}_{con}$.

(b) To enforce the attention consistency of invariance between different networks, the same input is fed to two branches, i.e., $\mathbf{x}' = \mathbf{x}$, and different networks with varied architecture and/or parameters, i.e., $f' \neq f$, are adopted for the two branches. In this case, the attention maps derived from the two branches are directly compared for the calculation of attention consistency loss $\mathcal{L}_{con}$.

These two kinds of attention consistency can also be combined by setting $\mathbf{x}' = T(\mathbf{x})$ and $f' \neq f$, with a unified attention consistency loss $\mathcal{L}_{con}$, in this two-branch framework. In each case, the classification loss $\mathcal{L}_{cls}$ and the attention consistency loss $\mathcal{L}_{con}$ are combined for the whole network training. During the testing, only one of

the branches is used for attribute recognition for computational efficiency and fair evaluation against existing methods.

### 5.2.2 Attribute Recognition and Visual Attention

Since human attribute recognition is an instance of multi-label visual recognition, the same cross entropy loss in Eq. (2.4) is adopted as the classification loss to train the network for attribute recognition. Meanwhile, the CAM-based interpretive attention maps are estimated according to Eq. (2.9).

Specially, let $\hat{\mathbf{y}} = f(\mathbf{x})$ and $\hat{\mathbf{y}}' = f'(\mathbf{x}')$ denote the output of two branches of the proposed framework, respectively. Accordingly, their classification losses can be defined as $\mathcal{L}_{cls}(\hat{\mathbf{y}}, \mathbf{y})$ and $\mathcal{L}_{cls}(\hat{\mathbf{y}}', \mathbf{y})$, respectively. For the same attribute $j$, the visual attention maps estimated from two branches can also be denoted as $h(\mathbf{x}, j, f)$ and $h(\mathbf{x}', j, f')$, respectively.

### 5.2.3 Visual Attention Consistency

To enforce the attention consistency, the comparing attention maps should first be aligned, i.e., elements of the same location in aligned attention maps correspond to the same location of the input image. Given an attribute $j$, let $\mathbf{Z}_j \in \mathbb{R}^{H \times W}$ and $\mathbf{Z}'_j \in \mathbb{R}^{H \times W}$ be the *aligned* attention maps computed from the two branches, respectively, the attention consistency loss is defined by

$$\mathcal{L}_{con}(\mathbf{Z}_j, \mathbf{Z}'_j) = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} |z_{jhw} - z'_{jhw}|^p, \tag{5.1}$$

where $z_{jhw}$ and $z'_{jhw}$ are the elements of the aligned attention maps $\mathbf{Z}_j$ and $\mathbf{Z}'_j$ at the location $(h, w)$, respectively, and $p > 0$ refers to a power term. Accordingly, the consistency loss on attention maps yields gradients for each attention pixel $z_{jhw}$ as

$$\frac{\partial \mathcal{L}_{con}(\mathbf{Z}_j, \mathbf{Z}'_j)}{\partial z_{jhw}} = \frac{p}{HW} |z_{jhw} - z'_{jhw}|^{p-1}. \tag{5.2}$$

Similarly, the gradients for the other branch can be calculated for each attention pixels $z'_{jhw}$. Equation (5.2) indicates the pixel-level local spatiality of the attributes is well considered by optimizing the proposed attention consistency loss. In the following, the construction of the aligned attention maps $\mathbf{Z}_j$ and $\mathbf{Z}'_j$ from the estimated CAM attention maps $h(\mathbf{x}, j, f)$ and $h(\mathbf{x}', j, f')$, respectively, is discussed to enforce the proposed two kinds of attention consistency.

**Attention Consistency 1 – Equivariance under Spatial Image Transforms:** When attention consistency of equivariance under spatial transforms is enforced, the inputs of two branches are the original image $\mathbf{x}$ and its transformed image $\mathbf{x}' = T(\mathbf{x})$, respectively, and the two branches use the same network, i.e, $f = f'$. The inverse transform $T^{-1}$ is conducted on the attention map estimated from the branch with the transformed image as input to make it spatially aligned with the attention map estimated from the branch with the original image as input, i.e.,

$$\begin{cases} \mathbf{Z}_j = h(\mathbf{x}, j, f), \\ \mathbf{Z}'_j = T^{-1}(h(\mathbf{T}(\mathbf{x}), j, f)). \end{cases} \tag{5.3}$$

Here $T$ is an image transform that does not change the visual perception, especially attention objects/contents for each attribute, in this image, such as image flipping, scaling, and rotation. While translation is also a typical spatial transform, its equivariance in both attention maps and final prediction has been well preserved in most existing deep networks, as verified in the later experiments.

**Attention Consistency 2 – Invariance between Different Networks:** When enforcing attention consistency of invariance between different networks, the same input, i.e., $\mathbf{x}' = \mathbf{x}$, is fed to two branches with different networks, i.e., $f' \neq f$, and the CAM attention maps estimated from two branches are already aligned and directly

comparable, i.e.,

$$\begin{cases} \mathbf{Z}_j = h(\mathbf{x}, j, f), \\ \mathbf{Z}'_j = h(\mathbf{x}, j, f'). \end{cases} \tag{5.4}$$

This way, two networks individually learn to recognize the same set of attributes and collaboratively learn attention maps for the same attribute from each other. Such a collaborative learning enables one network to learn missed knowledge that may be learned by the other network and vice versa, leading to enhanced learnings of both networks.

Note that the proposed method for attention consistency between networks is different from model ensemble [177], which trains multiple networks separately and then combines the predictions. In model ensemble, all the networks must be kept in both training and testing, resulting in significantly more parameters and computational consumption. Differently, the proposed method trains two networks simultaneously by achieving consistent attention maps and in the testing stage, we only deploy one individual network. This way, the proposed method for attention consistency of invariance between two different networks has increased computation consumption in training, but uses the same number of parameters and similar computation consumption as a single network in testing. In practice, a relatively shallower network can also be used for one of two branches to avoid introducing too many new parameters in training, with the goal of only deploying the other branch in testing.

**Combined Attention Consistency:** The combined attention consistency is also considered by enforcing both the equivariance under spatial transforms and the invariance between different networks. In this case, $\mathbf{x}' = T(\mathbf{x})$ and $f' \neq f$ are configured, i.e., the input of one branch is spatially transformed as the input of the other branch and two branches use different networks. As in Eq. (5.3), the inverse transform $T^{-1}$ need to be conducted on the attention map estimated from the branch with the trans-

formed image as input to make it spatially aligned with the attention map estimated from the branch with the original image as input, i.e.,

$$\begin{cases} \mathbf{Z}_j = h(\mathbf{x}, j, f), \\ \mathbf{Z}'_j = T^{-1}(h(\mathbf{T}(\mathbf{x}), j, f')). \end{cases} \tag{5.5}$$

This way, the unified loss Eq. (5.1) reflects a combination of the two kinds of consistency.

### 5.2.4 Consistency at Different Levels

This study proposes to apply consistency at the level of attention maps. Actually, consistency at other levels, such as the final prediction layer and certain feature layers, have been used in many previous works for improving deep network learning. For example, the widely used data augmentation strategy [70] assumes that a transformed image shares the same ground-truth classification as the original image and this can be regarded as enforcing the consistency of transform equivariance at the final step of recognition. Previous collaborative learning [172, 105, 109] also considers to enforce consistency between the final predictions of two different networks, as shown in Fig. 5.3(a). One previous work [167] considers the aggregated activations for information transfer between networks, which can be regarded as a feature-level consistency as shown in Fig. 5.3(c). The attention consistency proposed in this study is more attribute-specific on local regions, reflecting the local spatiality of attributes, as shown in Fig. 5.3(b).

More specifically, let's consider the use of prediction consistency loss of the $p$-th order in the previous collaborative learning, i.e.,

$$\mathcal{L}_{con}(\hat{y}_j, \hat{y}'_j) = |\hat{y}_j - \hat{y}'_j|^p. \tag{5.6}$$

Following Eqs. (2.8) and (2.9), we have

$$\hat{y}_j = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} z_{jhw} + b_j,$$

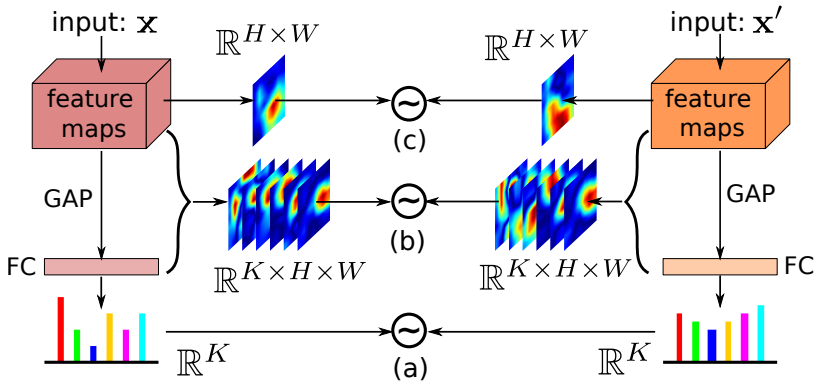


Figure 5.3 An illustration of consistency at different levels: (a) final prediction, (b) attention maps, and (c) feature aggregation.

and $\hat{y}'_j$ in a similar form. The produced gradients for an attention pixel would be

$$\begin{aligned}\frac{\partial \mathcal{L}_{con}(\hat{y}_j, \hat{y}'_j)}{\partial z_{jhw}} &= \frac{\partial \mathcal{L}_{con}(\hat{y}_j, \hat{y}'_j)}{\partial \hat{y}_j}\frac{\partial \hat{y}_j}{\partial z_{jhw}} \\ &= \frac{p}{HW}|\hat{y}_j - \hat{y}'_j|^{p-1}\end{aligned}. \tag{5.7}$$

The similar calculation can be applied to get the gradients at the pixel $z'_{jhw}$. Equation (5.7) clearly shows that, when using the prediction consistency, the gradients at different locations, i.e. with varying $(h, w)$, are the same for both networks and the local spatiality of each attribute is not well reflected in optimizing this loss. Besides, from Eq. (5.2) and Eq. (5.7), it can also be noticed that when setting $p = 1$, the proposed attention consistency degrades to the prediction consistency. Therefore, in the experiments, $p > 1$ is always configured. In the later Section 5.3, comparison experiments will be conducted by enforcing consistency at different levels to verify the effectiveness of the proposed methods.

### 5.2.5 End-to-end Training

Finally, the proposed two-branch network is learned in the end-to-end manner. The classification losses and the consistency loss are linearly combined by

$$\mathcal{L}_{total} = \mathcal{L}_{cls}(\hat{\mathbf{y}}, \mathbf{y}) + \mathcal{L}_{cls}(\hat{\mathbf{y}}', \mathbf{y}) + \lambda \mathcal{L}_{con}, \tag{5.8}$$

where $\lambda$ is a hyper-parameter to balance the two kinds of losses for each network learning. Classification losses $\mathcal{L}_{cls}(\hat{\mathbf{y}}, \mathbf{y})$ and $\mathcal{L}_{cls}(\hat{\mathbf{y}}', \mathbf{y})$ supervise the training of two branches, respectively, while the consistency loss is involved in the training of both branches.

## 5.3 Experiment

### 5.3.1 Datasets and Configurations

Experiments for this research is conducted on three representative human attribute datasets. **WIDER Attribute** [90] consists of images with complex scene contexts and 14 human attributes. The train-val set includes 28,345 samples (22,962 images in training set for model learning), while the test set includes 29,179 samples. **PA-100K** [97] contains 100,000 human bounding boxes in total, and has the largest number of training samples in the existing attribute datasets. 26 human attributes are annotated, and the training, validation and test sets are split with the ratio of 8:1:1. **RAP** [85] has a total number of 41,585 cropped human bounding boxes, with 69 attributes annotated, of which 51 attributes are usually recognized for evaluation. Among the existing human attribute datasets, RAP is the one with the largest number of attributes.

To be consistent with prior literatures, the experiments use the metric of mean Average Precision (mAP) on WIDER dataset, and the metrics of mean Accuracy (mA), instance Accuracy (Acc.), Precision (P), Recall (R) and F-1 score (F1)[1] on PA-100K and RAP datasets. Compared with mA, Acc., P, R, and F1, which rely on a specific threshold, e.g., 0.5, to produce binary prediction results, mAP reflects the model performance over all possible thresholds, leading to a more comprehensive evaluation on multiple attribute recognition.

[1]Detailed definitions can be found in [85].

On WIDER dataset, the exponential attribute loss weights of Eq. (2.6), initial learning rate of 0.001 (divided by 10 every 5 epochs), SGD optimizer and $p = 2$ in Eq. (5.1) are adopted. On PA-100K and RAP, input image size of $256 \times 128$, the attribute loss weights of Eq. (2.7), initial learning rate of 0.0001, $p = 3$ in Eq. (5.1), and Adam optimizer are adopted. The parameter $\lambda$ in Eq. (5.8) is set to 1 in our experiments. These configurations are mostly aligned with the following comparison methods.

Table 5.1 Performance comparison in terms of mean Average Precision (mAP, %) between the proposed methods and existing state-of-the-art methods on WIDER dataset. The baseline method is reproduced from the baseline of Da-HAR [165]. Attributes: 1 – Male, 2 – Long Hair, 3 – Sunglasses, 4 – Hat, 5 – T-shirt, 6 – Long Sleeves, 7 – Formal, 8 – Shorts, 9 – Jeans, 10 – Long Pants, 11 – Skirts, 12 – Face Mask, 13 – Logo, 14 – Plaid.

| Method | Backbone | Input size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|---|
| R*CNN ICCV'15 | VGG16 | $224 \times 224$ | 94 | 82 | 62 | 91 | 76 | 95 | 79 | 89 |
| DHC ECCV'16 | VGG16 | $224 \times 224$ | 94 | 82 | 64 | 92 | 78 | 95 | 80 | 90 |
| SRN CVPR'17 | ResNet101 | $224 \times 224$ | 95 | 87 | 72 | 92 | 82 | 95 | 84 | 92 |
| DIAA ECCV'18 | ResNet101 | $224 \times 224$ | 96 | 88 | 74 | 93 | 83 | 96 | 85 | 93 |
| Da-HAR AAAI'20 | ResNet101 | $256 \times 256$ | 97 | 89 | 76 | 96 | 85 | 97 | 86 | 92 |
| baseline | ResNet101 | $224 \times 224$ | 95 | 86 | 73 | 94 | 79 | 96 | 82 | 92 |
| VAC-TE (Ours) | ResNet101 | $224 \times 224$ | 96 | 89 | 76 | 96 | 83 | 97 | 85 | 94 |
| VAC-NI-A (Ours) | ResNet50 | $224 \times 224$ | 97 | 89 | 77 | 96 | 84 | 97 | 86 | 93 |
| VAC-NI-M (Ours) | ResNet101 | $224 \times 224$ | 97 | 90 | 78 | 96 | 84 | 97 | 86 | 93 |
| VAC-Combine (Ours) | ResNet101 | $224 \times 224$ | 97 | 90 | 79 | 96 | 85 | 98 | 86 | 94 |

| Method | Backbone | Input size | 9 | 10 | 11 | 12 | 13 | 14 | mAP |
|---|---|---|---|---|---|---|---|---|---|
| R*CNN ICCV'15 | VGG16 | $224 \times 224$ | 68 | 96 | 80 | 73 | 87 | 56 | 80.5 |
| DHC ECCV'16 | VGG16 | $224 \times 224$ | 69 | 96 | 81 | 76 | 88 | 55 | 81.3 |
| SRN CVPR'17 | ResNet101 | $224 \times 224$ | 80 | 96 | 84 | 76 | 90 | 66 | 85.1 |
| DIAA ECCV'18 | ResNet101 | $224 \times 224$ | 81 | 96 | 85 | 78 | 90 | 68 | 86.4 |
| Da-HAR AAAI'20 | ResNet101 | $256 \times 256$ | 81 | 97 | 87 | 79 | 91 | 70 | 87.3 |
| baseline | ResNet101 | $224 \times 224$ | 79 | 95 | 83 | 76 | 90 | 67 | 85.2 |
| VAC-TE (Ours) | ResNet101 | $224 \times 224$ | 83 | 96 | 87 | 79 | 92 | 69 | 87.5 |
| VAC-NI-A (Ours) | ResNet50 | $224 \times 224$ | 82 | 98 | 87 | 79 | 91 | 70 | 87.6 |
| VAC-NI-M (Ours) | ResNet101 | $224 \times 224$ | 83 | 98 | 88 | 80 | 92 | 71 | 88.1 |
| VAC-Combine (Ours) | ResNet101 | $224 \times 224$ | 84 | 98 | 88 | 80 | 92 | 71 | 88.4 |

### 5.3.2 Comparison with Existing Arts

#### Performance Comparison on WIDER Dataset

Firstly, experiments are conducted to compare the proposed methods with existing state-of-the-art approaches. On WIDER dataset, we compare with R*CNN [39], DHC [90], SRN [178], DIAA [124] and Da-HAR [165]. We can denote the proposed visual attention consistency of equivariance under spatial transforms and invaraince between differnt networks as VAC-TE (Transform Equivariant attention consistency) and VAC-NI (Network Invariant attention consistency), respectively. On WIDER dataset, we train VAC-TE by enforcing attention consistency of equivariance under a spatial transform randomly selected from *scaling* and *horizontal flipping*, with equal probability. For its scaling transform, we bi-linearly down-sample the image size from $224 \times 224$ to $192 \times 192$. Since many prior arts use ResNet101 as the backbone, we also adopt ResNet101 as the backbone of our methods for fair comparisons. In VAC-TE, both branches is constructed by an identical ResNet101 with shared parameters. In VAC-NI, one branch is constructed by ResNet101, which we regard as the *main branch*, denoted as VAC-NI-M, to learn for attribute recognition, and the other branch is constructed by ResNet50, which we regard as the *auxiliary branch*, denoted as VAC-NI-A. As mentioned above, for VAC-NI we mainly deploy/evaluate the main branch VAC-NI-M in the testing for computational efficiency and fair comparison with other single-network method. Using the relatively shallower auxiliary branch can help alleviate the increase of computation load in the training.

The performance comparison is reported in Table 5.1. Prior Da-HAR with ResNet101 as backbone achieves mAP of 87.3% over 14 human attributes. Our method VAC-TE achieves the mAP of 87.5%, while VAC-NI-M achieves the mAP of 88.1%. The comparison shows that considering attention consistency can improve the performance of human attribute recognition. Moreover, if we use the VAC-NI-A with ResNet50 for testing, the achieved performance is 87.6%, also outperforming the prior arts. As

discussed in Section 5.2.3, we can combine these two kinds of attention consistency into a unified consistency loss, where transform $T$ is randomly selected from scaling and flipping as mentioned above. As shown in Table 5.1, such combined consistency (VAC-Combine) can further improve the mAP of attribute recognition to 88.4%.

### Performance Comparison on PA-100K Dataset

Because prior arts use ResNet50 as the backbone on PA-100K evaluation, we follow the same protocol by taking ResNet50 as the backbone in the proposed methods. On PA-100K dataset, we train VAC-TE by enforcing attention consistency of equivariance under *horizontal flipping*. Table 5.2 shows the performance comparison between our method and prior methods, such as DeepMar [83], HPNet [97], VeSPA [125], PGDM [84], LGNet [95], ALM [143], JLPLS [142], and JLAC [141]. Based on ResNet50, we train a baseline model with an FC–BN (Batch Normalization) structure beside the FC layer for prediction regularization. Given the input size of $256 \times 128$, the baseline model achieves F1 score of 86.60%. When attention consistency of transform equivariance is enforced, VAC-TE achieves F1 score of 87.74%. Considering the attention consistency of invariance between networks, VAC-NI-M based on ResNet50 as the backbone achieves new state-of-the-art performance on F1 score of 88.23%, by using ResNet34 as the auxiliary branch. Furthermore, enforcing both kinds of attention consistency, VAC-Combine further improves the performance of F1 score to 88.41%, a new state-of-the-art performance on PA-100K dataset.

### Performance Comparison on RAP Dataset

On RAP dataset, the experiments use the same configurations as those applied to the experiments on PA-100K dataset. We also train VAC-TE by enforcing attention consistency of equivariance under *horizontal flipping*. The involved comparison methods include HPNet, VeSPA, PGDM, LGNet, JLPLS, CoCNN [45], JLAC, and

Table 5.2 Performance (%) comparison between our methods and prior methods on PA-100K.

| Method | | mA | Acc. | P | R | F1 |
|---|---|---|---|---|---|---|
| DeepMar ACPR'15 | | 72.70 | 70.39 | 82.24 | 80.42 | 81.32 |
| HPNet ICCV'17 | | 74.21 | 72.19 | 82.97 | 82.09 | 82.53 |
| VeSPA BMVC'17 | | 76.32 | 73.00 | 84.99 | 81.49 | 83.20 |
| PGDM ICME'18 | | 74.95 | 73.08 | 84.36 | 82.24 | 83.29 |
| LGNet BMVC'18 | | 76.96 | 75.55 | 86.99 | 83.17 | 85.04 |
| ALM ICCV'19 | | 80.68 | 77.08 | 84.21 | **88.84** | 86.46 |
| JLPLS TIP'19 | | 81.61 | 78.89 | 86.83 | 87.73 | 87.27 |
| JLAC AAAI'20 | | **82.31** | 79.47 | 87.45 | 87.77 | 87.61 |
| baseline-ResNet50 | | 81.58 | 78.97 | 86.32 | 86.89 | 86.60 |
| Ours | VAC-TE | 80.85 | 79.68 | 88.20 | 87.28 | 87.74 |
| | VAC-NI-M | 82.23 | 80.39 | 88.24 | 88.23 | 88.23 |
| | VAC-Combine | 82.19 | **80.66** | **88.72** | 88.10 | **88.41** |

Da-HAR. Similar to these methods, we conduct experiments on the five different train/test splits [85] and report the mean performance. As shown in Table 5.3, JLAC (ResNet50 backbone) and Da-HAR (ResNet101 backbone) achieve F1 scores of 80.82% and 80.72%, respectively. For fair comparison, we use the ResNet50 as our backbone. VAC-TE achieves F1 score of 80.79%, while VAC-NI-M with ResNet50 as the backbone achieves F1 score of 81.44%, of which the auxiliary branch adopts ResNet34 as the backbone. Also, when we enforce both kinds of attention consistency, VAC-Combine further improves the F1 score to 81.54%. While the mA performance of our proposed VAC-TE method is lower than that of JLAC, previous researches have pointed out that mean accuracy (mA) may not reflect the intrinsic dependency among multiple attributes and suffer from the imbalance issue between positive and negative samples of each human attribute.

Table 5.3 Performance (%) comparison between our methods and prior methods on RAP dataset.

| Method | mA | Acc. | P | R | F1 |
|---|---|---|---|---|---|
| HPNet ICCV'17 | 76.12 | 65.39 | 77.53 | 78.79 | 78.05 |
| VeSPA BMVC'17 | 77.70 | 67.35 | 79.51 | 79.67 | 79.59 |
| PGDM ICME'18 | 74.31 | 64.57 | 78.86 | 75.90 | 77.35 |
| LGNet BMVC'18 | 78.68 | 68.00 | 80.36 | 79.82 | 80.09 |
| JLPLS TIP'19 | 81.25 | 67.91 | 78.56 | 81.45 | 79.98 |
| CoCNN IJCAI'19 | 81.42 | 68.37 | 81.04 | 80.27 | 80.65 |
| JLAC AAAI'20 | **83.69** | 69.15 | 79.31 | **82.40** | 80.82 |
| Da-HAR AAAI'20 | 79.44 | 68.86 | 80.14 | 81.30 | 80.72 |
| baseline | 80.67 | 67.79 | 79.06 | 80.32 | 79.69 |
| VAC-TE (Ours) | 79.41 | 69.22 | 81.50 | 80.09 | 80.79 |
| VAC-NI-M (Ours) | 81.10 | 70.01 | 81.51 | 81.37 | 81.44 |
| VAC-Combine (Ours) | 81.30 | **70.12** | **81.56** | 81.51 | **81.54** |

### 5.3.3 Ablation Studies

In the following, we conduct ablations studies to further justify the detailed settings of the proposed methods, mainly on the WIDER dataset with input image size of $224 \times 224$.

#### Equivariance under Different Spatial Transforms

Different spatial transforms can be considered as $T$ of Eq. (5.3). Specifically, we focus on a set of frequently used transforms, including translation, rotation, scaling and flipping, for the ablation studies, since they do not change the visual perception of an image, i.e., the presence of human attributes. Certainly, in some extreme cases, e.g., down-sampling the input image to a very small size, the visual perception of an attribute may totally change. In this study, we choose appropriate parameters for these transforms to avoid such extreme cases. The four specific transforms we involve in this ablation study are 32-pixel translation to the right with zero-padding, 90° counter-clockwise rotation, bi-linear down-scaling from $224 \times 224$ to $192 \times 192$, and horizontal flipping.

As shown in Table 5.4, when there is no consideration of the attention consistency of transform equivariance in the network training for attribute recognition, the achieved mAP is 84.8% – it is slightly lower than the baseline performance of 85.2% in Table 5.1, because the latter also applies random horizontal flipping as data augmentation. When attention consistency of equivariance under either rotation, scaling or flipping is adopted for network regularization, the mAP performance for attribute recognition is improved. The combination of scaling and flipping (last row of Table 5.4), each with a random selection probability of 50%, leads to a further improved mAP performance of 87.5% and we chose this setting of transforms in the above comparison experiments against the state of the arts on the WIDER dataset. Since deep networks are inherently equivariant to translation, by using convolution and pooling operations, further enforcement of attention consistency of equivariance under translation does not introduce more performance improvement, as shown in Table 5.4.

Table 5.4 Performance (%) on WIDER Attribute dataset considering attention equivariance under different transforms, with ResNet101 as backbone. F1-C and F1-O [178] represent the macro and micro F1 scores evaluated by averaging per attribute results and on all images over all attributes, respectively.

| Transforms | mAP | F1-C | F1-O |
|---|---|---|---|
| Without | 84.8 | 75.5 | 80.6 |
| Translation | 84.6 | 75.3 | 80.1 |
| Rotation | 86.0 | 76.2 | 81.2 |
| Scaling | 86.5 | 76.5 | 81.6 |
| Flipping | 87.1 | 77.4 | 82.1 |
| Scaling & Flipping | **87.5** | **77.6** | **82.4** |

We also conduct an experiment to compare the attention consistency of equivariance under a spatial transform with using the same transform for data augmentation only. As shown in Table 5.5, enforcing attention consistency of equivariance under

certain transform achieve much better performance than using the same transform for data augmentation for network learning, except for the translation.

Table 5.5 Performance (%) on WIDER Attribute dataset using certain transform for data augmentation and attention consistency of equivariance, respectively. The backbone is ResNet50.

| Transform | Data Augmentation | | | Attention Consistency | | |
|---|---|---|---|---|---|---|
| | mAP | F1-C | F1-O | mAP | F1-C | F1-O |
| Without | 83.4 | 73.9 | 79.4 | – | – | – |
| Translation | 83.7 | 74.1 | 79.5 | 83.9 | 74.2 | 79.2 |
| Rotation | 83.2 | 73.2 | 78.5 | 85.0 | 75.1 | 80.2 |
| Scaling | 83.9 | 74.4 | 79.4 | 85.6 | 75.3 | 80.6 |
| Flipping | 84.2 | 74.6 | 80.0 | 86.3 | 76.4 | 81.2 |

### Consistency between Different Networks

In this section, we study the influence of using different auxiliary branches when enforcing the attention consistency between two networks. For this study, we take ResNet101 as the main branch in the proposed method, and consider ResNet50, ResNet152, DenseNet121, and DenseNet161 as the candidate backbone of the auxiliary branch. The comparison result in Table 5.6 shows that, even if the auxiliary branch itself, e.g. ResNet50 and DenseNet121, cannot achieve as good performance as the main branch, the main branch can still benefit from the proposed method by enforcing the attention consistency between the two branches. Moreover, when deeper networks, e.g., ResNet152 and DenseNet161, are used as the auxiliary branch, the attribute recognition performance of the main branch can be further improved, since deeper networks may provide more robust attention maps for collaborative attention learning. Table 5.6 also shows that our best performance of the main branch (ResNet101) on WIDER dataset is achieved by using ResNet152 as the auxiliary branch.

Table 5.6 Performance (mAP, %) of the main branch ResNet101 when the auxiliary branch using different backbones. Experiments are conducted on WIDER dataset with input size of $224 \times 224$.

| Auxiliary | VAC-NI-A | VAC-NI-M |
|---|---|---|
| Without | – | 85.2 |
| ResNet50 | 87.6 | 88.1 |
| ResNet152 | 88.6 | 88.4 |
| DenseNet121 | 87.5 | 88.3 |
| DenseNet161 | 88.4 | 88.3 |

### Quantitative Attention-Map Refinement

In this section, we conduct experiments to quantitatively examine whether the proposed attention consistency does improve the attention maps of the network. As mentioned earlier, constructing the ground-truth attention maps on a large-set of training images is very difficult for many attributes. Some attributes, such as "Age Between 18 and 60", may be related to ambiguous image regions and constructing its ground-truth attention map on an image may require a vision study involving a group of subjects following rigorous protocols. To quantitatively evaluate the quality of estimated attention maps, i.e., CAM, we select two attributes, "Long Hair" and "Shorts" in the WIDER dataset, with relatively unambiguous relevant regions, and manually annotate these regions. More specifically, we randomly select 200 test images for each of attributes "Long Hair" and "Shorts", and annotate the bounding boxes around the hair and shorts, respectively. By normalizing CAM attention maps to the value range of $[0, 1]$, we define an attention response ratio as the total attention values inside the bounding box over the area of bounding box. A higher attention response ratio indicates that the obtained attention map is more aligned with the annotated attention region and therefore, shows higher quality. Table 5.7 shows the results of two baseline methods, where ResNet50 and ResNet101 are trained without considering attention consistency, and the proposed methods enforcing attention

consistencies. For VAC-TE, we use ResNet101 with randomly selected scaling and flipping transforms as discussed above. For VAC-NI, we use ResNet101 as the main branch and ResNet50 as the auxiliary branch. Compared with the baselines, the proposed methods produce better attention maps by enforcing either type of attention consistency when recognizing these two attributes.

Table 5.7 Quantitative evaluation of the attention maps against the manually annotated attention regions for two attributes on selected test images in WIDER dataset. 'Baselines' indicates that networks are trained without considering attention consistency.

| Consistency | Nets | Attention Response Ratio (%) | |
|---|---|---|---|
| | | Long Hair | Shorts |
| Baselines | ResNet50 | 46.77 | 48.14 |
| | ResNet101 | 47.74 | 48.48 |
| VAC-TE | ResNet101 | 57.51 | 61.17 |
| VAC-NI-A | ResNet50 | 59.55 | 58.86 |
| VAC-NI-M | ResNet101 | 62.62 | 60.72 |

### Consistency at Different Levels

In this section, we conduct experiments on WIDER dataset to compare the use of consistency at different representation levels, including feature level, attention-map level and prediction level, as discussed in Section 5.2.4.

The result by enforcing the consistency of transform equivariance at different levels are reported in Table 5.8. The image transform adopted is horizontal flipping, and the backbone is ResNet50. Since the prediction for each attribute is a scalar without spatial information, we actually enforce flipping invariance of the prediction score for the prediction-level consistency. It can be regarded as an extension of the data augmentation, where the invariance is directly applied to the recognition result. For feature-level consistency, we enforce the feature equivariance of the last convolutional layer. For transform equivariance at each level, we use similar consistency loss by calculating element-wise difference, as in Eq. (5.1). As shown in Table 5.8, enforcing

Table 5.8 Performance (%) of enforcing flipping equivariance at different levels.

| Levels | w/o | Feature | Attention | Prediction |
|---|---|---|---|---|
| mAP(%) | 83.4 | 85.1 | 86.3 | 85.4 |

transform equivariance at the attention-map level achieves the best results, since the local spatiality of attribute recognition is well embedded. Also, compared with feature equivariance under transforms, attention equivariance under the same transforms encodes attribute specific spatial information in the network learning, leading to better performance.

For the consistency between networks, we also compare the performance by enforcing attention consistency against the feature/prediction-level consistency, as discussed in Section 5.2.4. As shown in Table 5.9, both the considerations of feature-level consistency (Fig. 5.3(c)) and prediction-level consistency (Fig. 5.3(a)) can improve the attribute recognition performance. But the proposed method achieves the largest improvement by considering the attention-level consistency between two networks. Both experiments in Table 5.8 and Table 5.9 demonstrate that attention-level consistency is superior to feature- and prediction-level consistency for human attribute recognition.

Table 5.9 Performance comparison (mAP(%)) of using different-level consistency for collaborative learning on WIDER dataset. Two networks are ResNet50 and ResNet101, and the input size is $224 \times 224$.

| Levels | ResNet50 | ResNet101 |
|---|---|---|
| w/o | 84.3 | 85.2 |
| Feature ICLR'17 | 85.7 | 86.5 |
| Prediction CVPR'18 | 86.8 | 87.6 |
| Attention (Ours) | 87.6 | 88.1 |

### Comparison to Network Ensemble

In the above Section 5.2.3, we discuss the difference between the proposed method by enforcing attention consistency between networks and prior works on model ensemble [177], which integrate predictions from multiple networks in the testing. Since our proposed method only deploys one branch, it has much fewer parameters and takes much less computation time than model-ensemble methods in the testing. In this section, we conduct an experiment to compare the performance of the proposed method and the model-ensemble method. For simplicity, we average the predictions from two networks for model ensemble. As shown in Table 5.10, when two networks, e.g., ResNet50 and ResNet101, are trained separately, i.e., "w/o VAC", the model ensemble achieves better performance than each of them. When two networks are collaboratively learned by enforcing attention consistency by using our proposed method, either branch of our collaboratively trained networks performs better than the direct model ensemble. Moreover, we can also average the predictions from the trained two branches of the proposed method, which leads to further performance improvement. These results verify not only the effectiveness of the proposed attention consistency between networks, but also its complementarity to ensemble methods.

### Hyper-parameter Influence

Based on the consistency loss between two networks, we conduct experiments to investigate the influence of the power term $p$ and show the recognition performance on two datasets in Fig. 5.4. Specifically, on WIDER dataset, we use ResNet101 as the main branch and ResNet50 as the auxiliary branch, while on PA-100K dataset, we use ResNet50 as the main branch and ResNet34 as the auxiliary branch. As shown in Fig. 5.4(a), the best mAP performance on WIDER dataset (88.1%) is achieved by using $p = 2$, while an overly large power, e.g., $p = 4$, makes the consistency loss dominate the network learning, leading to reduced mAP performance. On PA-100K,

Table 5.10 Performance comparison between the proposed method and model ensemble. 'VAC' indicates attention consistency between networks.

| Datasets | Networks | w/o VAC | with VAC |
|---|---|---|---|
| WIDER (mAP, %) | ResNet50 | 84.3 | 87.6 |
| | ResNet101 | 85.2 | 88.1 |
| | Ensemble (50 & 101) | 86.6 | 88.3 |
| PA-100K (mAP, %) | ResNet34 | 70.91 | 74.07 |
| | ResNet50 | 71.03 | 74.32 |
| | Ensemble (34 & 50) | 73.46 | 74.97 |
| PA-100K (F1, %) | ResNet34 | 86.10 | 88.12 |
| | ResNet50 | 86.60 | 88.23 |
| | Ensemble (34 & 50) | 87.83 | 88.35 |
| RAP (F1, %) | ResNet34 | 78.98 | 81.02 |
| | ResNet50 | 79.69 | 81.44 |
| | Ensemble (34 & 50) | 80.71 | 81.65 |

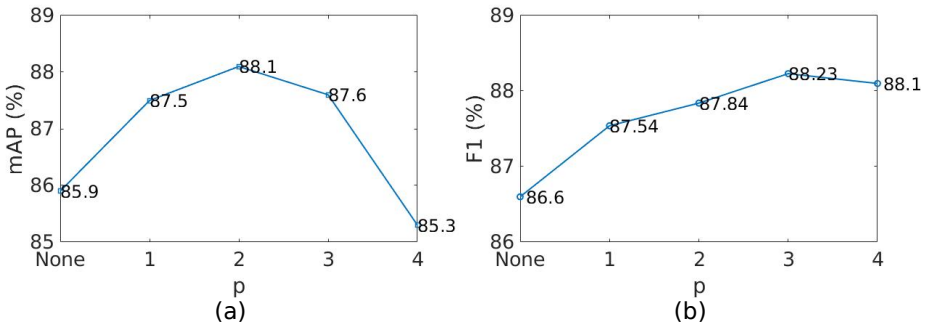


Figure 5.4 Performance of attribute recognition by setting different values for $p$ in the attention consistency between two networks.

there exists more severe data imbalance. A larger power term $p$ in the attention consistency loss is desired to emphasize the pixel-wise difference between attention maps for the same attribute. As shown in Fig. 5.4(b), the best F1 performance of attribute recognition on PA-100K is achieved by setting $p = 3$.

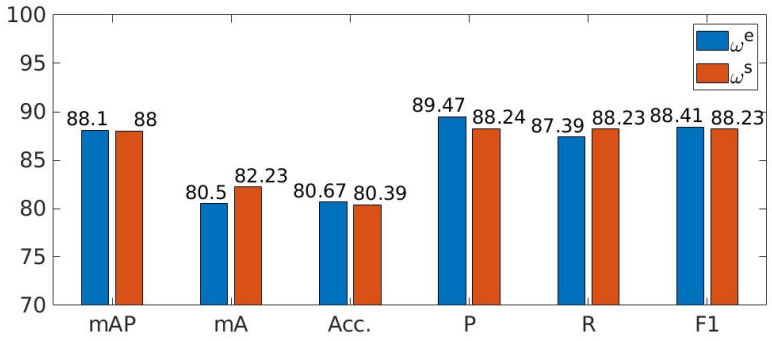


Figure 5.5 Attribute recognition performance (mAP, %) by using different attribute weights in the classification loss on WIDER dataset, and mA, Acc. P, R, and F1 are reported on PA-100K.

#### Attribute Weights in Classification Loss

We adopt two attribute weighting strategies, as shown in Eq. (2.6) and Eq. (2.7), respectively, on different datasets to fairly compare our method with prior works. We further compare the strategies on the same dataset in Fig. 5.5. The experiment results further demonstrate that the proposed method is actually robust to both weighting strategies.

### 5.3.4 Qualitative Analysis

To qualitatively analyze the proposed method for attribute recognition, we visually compare the attention maps from the baseline ResNet101 without using attention consistency, and those enhanced with two kinds of attention consistency. As shown in Fig. 5.6, each row illustrates the attention maps for recognizing an attribute from the same image by different methods. Attention maps estimated by the baseline method without enforcing attention consistency may highlight visually irrelevant regions for certain attribute recognition, e.g., leg regions in recognizing the attribute "T-shirts" in the second column of row (a). When attention consistency of transform equivariance is adopted, the attention map is refined in the third column of row (a) by paying

more attention on upper body. Furthermore, enforcing attention consistency between networks can also refine the attention maps by focusing attention on upper body as shown in the fourth and fifth columns of row (a) with ResNet50 and ResNet101 as backbones, respectively.

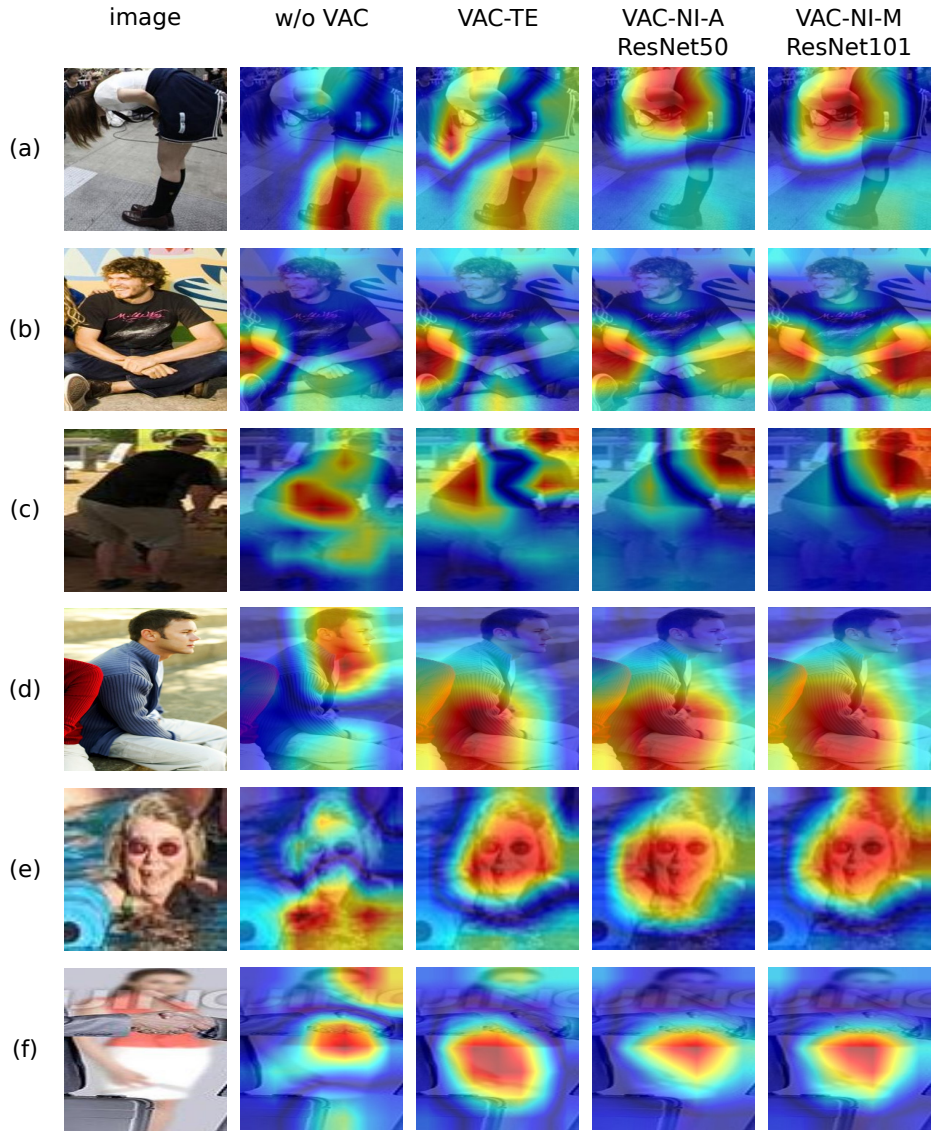


Figure 5.6 Qualitative comparison of attention maps estimated in recognizing the same attribute (each row) by using different methods. The attributes to be recognized in each row are (a) T-shirt, (b)Jeans, (c) Hat (d) Long Pants, (e) Long Hair and (f) Skirt.

Moreover, for recognizing attributes "T-shirt", "Jeans" and "Hat", VAC-TE refines the corresponding attention maps, but may still miss/highlight some relevant regions/irrelevant regions for the attribute, e.g., highlighted leg regions for "T-shirt", missed left-leg regions for "Jeans" and highlighted waist regions for "Hat" in the third column of rows (a), (b) and (c), respectively. By contrast, VAC-NI-A and VAC-NI-M better highlight upper body, two legs and head regions for recognizing "T-shirt", "Jeans" and "Hat", respectively. This is aligned with the quantitative results which also show that VAC-NI achieves better performance than VAC-TE. For recognizing attributes "Long Pants", "Long Hair" and "Skirt" in rows (d), (e) and (f), respectively, enforcing either kind of attention consistency makes the corresponding attention maps to better highlight the correct image regions, e.g., leg, head and lower body. These qualitative results verify that the proposed two kinds of attention consistency can refine the visual attention map of networks in recognizing human attributes.

## 5.4 Chapter Summary

This study proposed new methods to improve the plausibility of deep network attention maps to improve the performance of human attribute recognition. Specifically, we designed a two-branch framework to enforce the attention consistency during network learning for attribute recognition. In this framework, we formulated two kinds of attention consistency, i.e., equivariance under spatial transforms and invariance between different networks, and defined corresponding attention consistency losses, which are combined with the initial classification loss for network learning. We conducted comprehensive experiments on three representative datasets for human attribute recognition and verified the effectiveness of enforcing attention consistency for attribute recognition by achieving new state-of-the-art performances on all these datasets.

# Chapter 6

# Collaborative Learning on Biased Distributions for Long-tailed Label Distribution

Long-tailed data distribution is a common label imbalance issue in many practical multi-label visual recognition tasks and the direct use of these data, i.e., uniform sampling, for training usually leads to relatively low performance on tail classes. While re-balanced data sampling can improve the performance on tail classes, it may also hurt the performance on head classes in training due to label co-occurrence. Thus, deep network training from either uniform sampling or re-balanced sampling of the long-tailed data for multi-label visual recognition is actually learning from a biased distribution. Improving the performance of recognizing a sub-group classes is at the expense of decreasing the performance of recognizing another sub-group of classes.

This study proposes a new approach to train on both uniform and re-balanced samplings in a collaborative way, resulting in performance improvement on both head and tail classes. More specifically, we design a visual recognition network with two branches: one takes the uniform sampling as input while the other takes the re-balanced sampling as the input. For each branch, we conduct visual recognition using a binary-cross-entropy-based classification loss with learnable logit compensation. We further define a new cross-branch loss to enforce the consistency when the same input image goes through the two branches. We conduct extensive experiments on VOC-LT and COCO-LT datasets. The results show that the proposed method

significantly outperforms previous state-of-the-art methods on long-tailed multi-label visual recognition.

To summarize, the main contributions of this work are:

1 We propose the use of both uniform and re-balanced samplings of the same training set for long-tailed multi-label visual recognition.

2 We develop a two-branch network, as well as a cross-branch loss to enforce the consistency between two branches, for collaborative learning on both uniform and re-balanced samplings.

3 We conduct extensive experiments on VOC-LT and COCO-LT datasets to verify that the proposed method can simultaneously improve the performance of both head and tail classes.

## 6.1 Overview

Re-balanced data sampling [10, 129, 6, 47] is a proven effective approach for addressing the long-tailed visual recognition. It achieves class-wise balance by either down-sampling the head-class data or up-sampling the tail-class data. However, while re-balanced sampling can improve the recognition performance of tail classes, it may simultaneously decrease the performance of some head classes due to label co-occurrence in multi-label recognition [166]. Since performance of different classes, either head or tail ones, is usually considered to be equally important in multi-label visual recognition, this study develops a new method that can combine different data samplings for improving the performance of both head and tail classes.

We consider the uniform and re-balanced samplings yielding two biased data distributions for long-tailed multi-label visual recognition. Given a long-tailed training set for multi-label recognition, the uniform sampling leads to the original long-tailed distribution bias towards head classes, while the re-balanced sampling yields another

distribution bias towards tail classes. Our basic idea is to use each of them to train a branch of a two-branch network, where two branches follow the same architecture. We further define a loss that enforces the consistency across the two branches for the same input to achieve a collaborative training, inspired by the previous mutual learning [172] and co-regularization [105]. The cross-branch consistency can compromise two branches to make the deep network learn from a relatively more balanced distribution somewhere between these two biased distributions, so that recognition performance of all classes is improved.

More specifically, as shown in Fig. 6.1(b), the two branches have the same architecture but different parameters to reflect the different distributions of their respective inputs. For each branch, the binary-cross-entropy-based multi-label classification loss with learnable logit compensation is defined for multi-label visual recognition. For combining two branches, we introduce another loss to collaboratively enforce the prediction consistency across the two branches when the same input image is fed to the two branches. Finally, this two-branch network is trained in an end-to-end manner by minimizing both classification and consistency losses. During the test phase, each test image is fed to both branches without considering cross-branch paths and the average of predictions from the two branches is taken as the final prediction.

Different from previous mutual learning methods [172, 105], where the two branches always take the input from a single distribution, as shown in Fig. 6.1(a), the proposed method learns two branches from different inputs generated by different samplings and the same input for two branches is only used for computing the consistency loss.

## 6.2 Methodology

Similar to notations denoted as in Section 2.3, let the training set for the long-tailed multi-label visual recognition (LTML) be $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ are the $N$ training images and $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N\}$ are their respective ground-truth

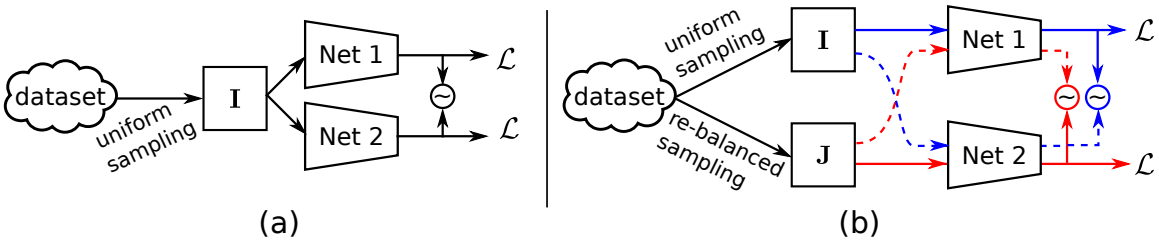


Figure 6.1 An illustration of the difference between (a) the previous mutual learning [172]/co-regularization [105] networks, where the input from the same distribution is always fed to the two branches, and (b) the proposed network where different inputs, from different samplings, are fed to the two branches. We only use the same input for the two branches for computing the consistency loss. **I** and **J** are mini-batch images, $\sim$ indicates the consistency measurement, and $\mathcal{L}$ is the classification loss.

class labels. Specifically, each $\mathbf{y}_i = [y_{i1}, y_{i2}, \cdots, y_{iK}], i = 1, 2, \cdots, N$ is a binary $K$-dimensional vector where $y_{ik} = 1$ indicates the presence of label $k$ in image $i$ and $y_{ik} = 0$ otherwise, with $k = 1, 2, \cdots, K$. $K$ is the total number of labels for the visual recognition. There may be multiple elements of value 1 in each $\mathbf{y}_i$ for multi-label visual recognition.

### 6.2.1 Framework Overview

Given that $(\mathbf{X}, \mathbf{Y})$ follows a long-tail distribution in terms of class labels, we use both uniform and re-balanced samplings in preparing the inputs for network training. For the uniform sampling, each image $\mathbf{x}_i \in \mathbf{X}$ is sampled with an instance-level probability of $1/N$. For the re-balanced sampling [129, 65, 166], images of each class are sampled with a class-level probability of $\frac{1}{K}$, and thus, each image $\mathbf{x}_i$ is sampled with a probability of $\frac{1}{K}\sum_{k=1}^{K}\frac{y_{ik}}{N_k}$, where $N_k$ is the number of images with class label $k$ in the training set. By sampling the original training set $M$ times, the re-balanced sampling actually provides us a new relatively class-balanced training set $(\mathbf{X}', \mathbf{Y}')$, with $M$ samples.

As shown in Fig. 6.2, the two branches of the proposed network share the same bottom network $\varphi$, followed by another CNN module, denoted as 'Subnet-U' in the branch for the uniform sampling and 'Subnet-R' in the branch for the re-balanced sampling. Subnet-U and Subnet-R have the same architecture but trained with different parameters, as shown in Fig. 6.2. To be specific, the shared bottom network is the conventional ResNet [49] excluding the last stage. For Subnet-U and Subnet-R, we first include an identical copy of the last stage of ResNet, as shown by $f_1$ and $g_1$ in Fig. 6.2. After that, a linear classifier in the form of a fully connected layer is added to each branch, as shown by $f_2$ and $g_2$ in Fig. 6.2 for multi-label recognition. When feeding images $\mathbf{x}_i^u \in \mathbf{X}$ and $\mathbf{x}_j^r \in \mathbf{X}'$ to the two branches respectively, we obtain $K$-dimensional logits for the two branches as

$$\begin{cases} \mathbf{u}_i = f_2(f_1(\varphi(\mathbf{x}_i^u))), \\ \mathbf{r}_j = g_2(g_1(\varphi(\mathbf{x}_j^r))). \end{cases} \tag{6.1}$$

By formulating the task as multiple binary image classifications, we apply logistic linear regression on logits $\mathbf{u}_i \in \mathbb{R}^K$ and $\mathbf{r}_j \in \mathbb{R}^K$ to learn the two branches, respectively. The solid arrows in blue and red in Fig. 6.2 indicate the classification paths for the two branches, respectively. The binary-cross-entropy-based classification losses $\mathcal{L}_{cls}(\mathbf{u}_i, \mathbf{y}_i^u)$ and $\mathcal{L}_{cls}(\mathbf{r}_j, \mathbf{y}_j^r)$ are adopted for respective branch optimization, where $(\mathbf{u}_i, \mathbf{y}_i^u)$ and $(\mathbf{r}_j, \mathbf{y}_j^r)$ represent the pair of predicted logits and ground-truth labels for the $i$-th image in $\mathbf{X}$ and the $j$-th image in $\mathbf{X}'$, respectively.

We further cross the inputs of two branches and estimate the logits, indicated by the blue/red dashed arrows in Fig. 6.2 and obtain

$$\begin{cases} \hat{\mathbf{u}}_i = g_2(g_1(\varphi(\mathbf{x}_i^u))), \\ \hat{\mathbf{r}}_j = f_2(f_1(\varphi(\mathbf{x}_j^r))). \end{cases} \tag{6.2}$$

To enforce the two branches to make consistent predictions from the same input, we introduce a mean-square-error based consistency loss $\mathcal{L}_{con}(\mathbf{u}_i, \hat{\mathbf{u}}_i)$ and $\mathcal{L}_{con}(\mathbf{r}_j, \hat{\mathbf{r}}_j)$

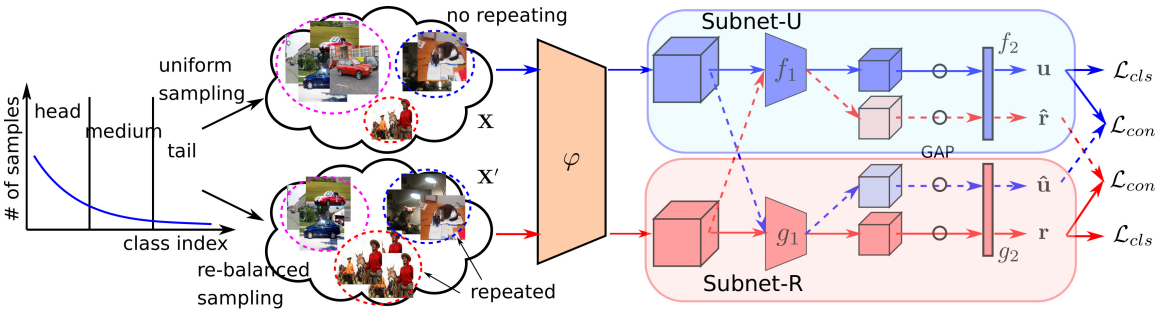


Figure 6.2 An illustration of the proposed network for long-tailed multi-label visual recognition. GAP denotes the global average pooling.

between the logits from different branches, indicate by the same color arrows (one dashed and one solid) in Fig. 6.2.

Finally, the network is learned by jointly minimizing the loss function

$$\begin{aligned}\mathcal{L}(\mathbf{x}_i^u, \mathbf{x}_j^r; \mathbf{y}_i^u, \mathbf{y}_j^r) =& \mathcal{L}_{cls}(\mathbf{u}_i, \mathbf{y}_i^u) + \mathcal{L}_{cls}(\mathbf{r}_j, \mathbf{y}_j^r) + \\ & \lambda(\mathcal{L}_{con}(\mathbf{u}_i, \hat{\mathbf{u}}_i) + \mathcal{L}_{con}(\mathbf{r}_j, \hat{\mathbf{r}}_j)),\end{aligned} \tag{6.3}$$

where $(\mathbf{x}_i^u, \mathbf{y}_i^u) \in (\mathbf{X}, \mathbf{Y})$, $(\mathbf{x}_j^r, \mathbf{y}_j^r) \in (\mathbf{X}', \mathbf{Y}')$, and $\lambda$ is a hyper-parameter to balance the two kinds of loss functions.

### 6.2.2 Conventional Classification Loss

Conventionally, the weighted sigmoid cross entropy loss [83, 42, 141] is used for multi-label visual recognition, in the form of multiple binary image classifications, as discussed in Eq. (2.4) in Section 2.3. Taking the branch for the uniform sampling as an example, this loss is

$$\begin{aligned}\mathcal{L}_{cls}(\mathbf{u}_i, \mathbf{y}_i^u) = -\frac{1}{K}\sum_{k=1}^{K} \omega_k \, & (y_{ik}^u \log(\varsigma(u_{ik})) + \\ & (1 - y_{ik}^u) \log(1 - \varsigma(u_{ik}))) \, ,\end{aligned} \tag{6.4}$$

where $u_{ik}$ and $y_{ik}^u$ are the $k$-th elements of the predicted logits $\mathbf{u}_i$ and the ground-truth label $\mathbf{y}_{ik}^u$, respectively, corresponding to the $k$-th label. Besides, $\omega_k = y_{ik}^u e^{1-\rho} + (1 - y_{ik}^u)e^{\rho}$ is the loss weight for the $k$-th label, depending on its ratio of positive samples

$\rho = N_k/N$, and $\varsigma$ is the sigmoid function converting logits in $\mathbb{R}$ to probabilities in the range of [0, 1] by

$$\varsigma(u_{ik}) = 1/(1 + e^{-u_{ik}}). \tag{6.5}$$

The classification loss $\mathcal{L}_{cls}(\mathbf{r}_j, \mathbf{y}_j^r)$ for the other branch can be defined in the same way.

### 6.2.3 Logit Compensation

As discussed in [8, 166], when using the weighted sigmoid cross entropy loss for classification, the imbalance between the numbers of positive and negative samples in each class could push their unbounded logit values away from zero with different distances, leading to class-specific over-fitting. In this section, we address this issue by further compensating the logits of positive and negative samples, respectively.

For simplicity, we assume that logit output of the network for each label recognition conforms to a normal distribution. Suppose the logit for positive samples of the $k$-th label conforms to a normal distribution with mean $\mu_k^p$ and standard deviation $\sigma_k^p$, and the logit for negative samples of the same label conforms to a normal distribution with mean $\mu_k^n$ and standard deviation $\sigma_k^n$. The mean logit values $\{\mu_1^p, \mu_2^p, \cdots, \mu_K^p\}$ and $\{\mu_1^n, \mu_2^n, \cdots, \mu_K^n\}$, and standard deviations $\{\sigma_1^p, \sigma_2^p, \cdots, \sigma_K^p\}$ and $\{\sigma_1^n, \sigma_2^n, \cdots, \sigma_K^n\}$ are then used to compensate the logits before feeding to the classification loss in Eq. (6.4). Thus, the classification loss (6.4) is upgraded to

$$\begin{aligned}\mathcal{L}_{cls}(\mathbf{u}_i, \mathbf{y}_i^u) = -\frac{1}{K}\sum_{k=1}^{K} \omega_k \big(& y_{ik}^u \log(\varsigma(u_{ik} \cdot \sigma_k^p + \mu_k^p)) \\ & + (1 - y_{ik}^u) \log(1 - \varsigma(u_{ik} \cdot \sigma_k^n + \mu_k^n))\big)\,.\end{aligned} \tag{6.6}$$

The classification loss $\mathcal{L}_{cls}(\mathbf{r}_j, \mathbf{y}_j^r)$ is upgraded with logit compensation in the same way. All the above means and standard deviations are learnable parameters. Compared with previous logit-adjustment methods [8, 166], this simple compensation does not introduce additional empirical hyper-parameters that require manually tuning.

### 6.2.4 Logit Consistency between Branches

In the ideal case, when we feed the same input image to the two branches, the output predictions shall approximate the ground-truth labels with the network optimizations. However, since the two branches attempt to fit the differently biased distributions of input data, they may produce different prediction results with the same input, e.g., the two branches may show different recognition performance. As mentioned above, we define a cross-branch consistency loss based on the mean square error of logits computed from the same input image but through different branches. Taking the input from the uniform sampling as an example, this loss is

$$\mathcal{L}_{con}(\mathbf{u}_i, \hat{\mathbf{u}}_i) = \frac{1}{K}\sum_{k=1}^{K}(u_{ik} - \hat{u}_{ik})^2, \tag{6.7}$$

where $u_{ik}$ and $\hat{u}_{ik}$ are the $k$-th elements of $\mathbf{u}_i$ and $\hat{\mathbf{u}}_i$, respectively. For the input from the re-balanced sampling, the consistency loss $\mathcal{L}_{con}(\mathbf{r}_j, \hat{\mathbf{r}}_j)$ can be defined in the same way.

Different from existing works on collaborative training [172, 105], which define consistency on probabilities, e.g., softmax/sigmoid outputs, for visual recognition, here we measure the consistency between logits of different branches from the same input. In training multi-label classifiers, due to the sigmoid normalization in Eq. (6.5), gradients could vanish on highly confident probabilities. For example, when the consistency loss is applied to probabilities, we have loss $\mathcal{L}_{con}(\varsigma(\mathbf{u}_i), \varsigma(\hat{\mathbf{u}}_i))$ and the gradients propagated to the logits $\mathbf{u}_i$ would be:

$$\begin{aligned}\frac{\partial \mathcal{L}_{con}(\varsigma(\mathbf{u}_i), \varsigma(\hat{\mathbf{u}}_i))}{\partial \mathbf{u}_i} &= \frac{\partial \mathcal{L}_{con}(\varsigma(\mathbf{u}_i), \varsigma(\hat{\mathbf{u}}_i))}{\partial \varsigma(\mathbf{u}_i)}\frac{\partial \varsigma(\mathbf{u}_i)}{\partial \mathbf{u}_i} \\ &= \frac{\partial \mathcal{L}_{con}(\varsigma(\mathbf{u}_i), \varsigma(\hat{\mathbf{u}}_i))}{\partial \varsigma(\mathbf{u}_i)}\varsigma(\mathbf{u}_i)(1 - \varsigma(\mathbf{u}_i)).\end{aligned} \tag{6.8}$$

If the predicted probabilities are highly confident, e.g. $\varsigma(\mathbf{u}_i) \simeq 1$ or $\varsigma(\mathbf{u}_i) \simeq 0$, the gradients from consistency loss are close to zero. Differently, we define the consistency

loss based on logits, with which the gradients propagated to the logits $\mathbf{u}_i$ would be:

$$\frac{\partial \mathcal{L}_{con}(\mathbf{u}_i, \hat{\mathbf{u}}_i)}{\partial \mathbf{u}_i} = \frac{2}{K}(\mathbf{u}_i - \hat{\mathbf{u}}_i). \tag{6.9}$$

We can see that these gradients do not have the above gradient vanishing issue under high-confident predictions.

### 6.2.5 Model Inference

To conduct model inference on test images, we simply feed all the test images to both branches of the trained network one by one. The paths following the dashed arrows in Fig. 6.2 are not used. For each input test image, the predictions of two branches are averaged as the final prediction result.

## 6.3 Experiment

### 6.3.1 Datasets and Configurations

As in [166], we conduct experiments on two datasets for long-tailed multi-label visual recognition: VOC-LT and COCO-LT. They are artificially constructed from two multi-label visual recognition benchmarks, VOC [29] and MS-COCO [92], respectively.

**VOC-LT** is sampled from the 2012 train-val set of VOC [29] based on a Pareto distribution as described in [98]. The training set contains 1,142 images and 20 class labels, and the number of images per class ranges from 4 to 775. The 20 classes are split into three groups according to the number of training samples per class: a head class has more than 100 samples, a medium class has 20 to 100 samples, and a tail class has less than 20 samples. The ratio of head, medium and tail classes after such splitting is 6:6:8. The testing set is constructed on the 2007 test set of VOC, with 4,952 images.

**COCO-LT** is created from the 2017 version of MS-COCO [92] by following a similar way. The training set of this long-tailed dataset contains 1,909 images and 80 class labels, and the number of images per class ranges from 6 to 1,128. The ratio of head, medium and tail classes is 22:33:25, following a similar split as in VOC-LT. The test set consists of all 5,000 images in the test set of MS-COCO-2017.

**Configurations:** Following [166] and the conventional multi-label visual recognition [178, 160, 42], we use the mean Average Precision (mAP) to evaluate the performance of long-tailed multi-label visual recognition. We use the similar configurations as in [166] in our experiments for a fair comparison with this prior state-of-the-art method. Specifically, we use the ResNet50 [47] pre-trained on ImageNet [70, 122] as the backbone and input images are resized to the spatial dimension of $224 \times 224$. The standard data augmentations are applied as in [166]. The SGD with momentum of 0.9 and weight decay of 0.0001 is adopted as the optimizer. The hyper-parameter $\lambda$ in Eq. (6.3) is set to 0.1 constantly. In the classification loss with logit compensation in Eq. (6.6), the mean values are initialized to 0, while the standard deviations are initialized to 1. The initial learning rate is set to 0.01. All experiments are conducted on PyTorch 1.4.0.

### 6.3.2 Comparison with Prior Arts

First of all, to verify the effectiveness of the proposed method, we compare the mAP performance between our method and previous methods on both long-tailed datasets. The comparison methods include Empirical Risk Minimization (ERM), conventional Re-Weighting (RW) using the inverse proportion to the square root of class frequency, Re-Sampling (RS) [129], Focal Loss [91], ML-GCN [12], OLTR [98], LDAM [8], CB Focal [19], BBN [176] and DB Focal [166]. The mAP performance of different methods are shown in Table 6.1. The prior best performance is achieved by DB Focal [166] – mAP of 78.94% over all classes on VOC-LT and 53.55% over all classes on COCO-LT.

Table 6.1 mAP performance of the proposed method and comparison methods. The notation * indicates the reproduced results based on our experiment environment. Other comparison results are taken from [166].

| Datasets | VOC-LT | | | | COCO-LT | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | total | head | medium | tail | total | head | medium | tail |
| ERM | 70.86 | 68.91 | 80.20 | 65.31 | 41.27 | 48.48 | 49.06 | 24.25 |
| RW | 74.70 | 67.58 | 82.81 | 73.96 | 42.27 | 48.62 | 45.80 | 32.02 |
| Focal Loss [91] | 73.88 | 69.41 | 81.43 | 71.56 | 49.46 | 49.80 | 54.77 | 42.14 |
| RS [129] | 75.38 | 70.95 | 82.94 | 73.05 | 46.97 | 47.58 | 50.55 | 41.70 |
| ML-GCN [12] | 68.92 | 70.14 | 76.41 | 62.39 | 44.24 | 44.04 | 48.36 | 38.96 |
| OLTR [98] | 71.02 | 70.31 | 79.80 | 64.95 | 45.83 | 47.45 | 50.63 | 38.05 |
| LDAM [8] | 70.73 | 68.73 | 80.38 | 69.09 | 40.53 | 48.77 | 48.38 | 22.92 |
| CB Focal [19] | 75.24 | 70.30 | 83.53 | 72.74 | 49.06 | 47.91 | 53.01 | 44.85 |
| BBN* [176] | 73.37 | 71.31 | 81.76 | 68.62 | 50.00 | 49.79 | 53.99 | 44.91 |
| DB Focal [166] | 78.94 | 73.22 | 84.18 | 79.30 | 53.55 | 51.13 | 57.05 | 51.06 |
| DB Focal* [166] | 78.42 | 74.13 | 83.19 | 78.06 | 54.33 | 50.06 | 57.22 | 54.27 |
| baseline-uniform | 77.15 | 73.14 | 83.49 | 75.41 | 53.15 | 51.61 | 57.17 | 49.21 |
| baseline-re-balanced | 78.36 | 71.72 | 83.58 | 79.41 | 52.76 | 48.67 | 56.87 | 50.94 |
| Ours | **81.44** | **75.68** | **85.53** | **82.69** | **56.90** | **54.13** | **60.59** | **54.47** |

We further reproduce DB Focal, denoted as DB Focal* in Table 6.1, on our platform based on its implementation [1] and achieve similar mAP performances as the ones reported in [166].

We train two baselines for the proposed method with the conventional classification loss and different samplings. Specifically, we train the proposed network only with one branch using the uniform sampling and re-balanced sampling, respectively, with the weighted classification loss in Eq. (6.4). This way, we obtain two baselines: baseline-uniform and baseline-re-balanced, respectively. From Table 6.1, we can see that both baselines achieve lower mAP performance than DB Focal (or DB Focal*) – mAP performances of two baselines on VOC-LT are 77.15% and 78.36%, respectively, and those on COCO-LT are 53.15% and 52.76%, respectively. The proposed method can significantly increase the mAP performance on both datasets: mAP performance is improved to 81.44% on VOC-LT (increased by 3.02% from DB Focal*)

[1] https://github.com/wutong16/DistributionBalancedLoss

and to 56.90% on COCO-LT (increased by 2.63% from DB Focal*). Besides, the proposed method also achieves the new state-of-the-art mAP performance for both head, medium and tail classes on both datasets.

### 6.3.3 Quantitative Analysis

#### Ablation Analysis

To further analyze how the proposed method improves mAP performance for long-tailed multi-label recognition, we conduct a set of ablation studies and report the results in Table 6.2. We first conduct an experiment by using a simple branch-ensemble method which averages the predictions from the two branches as the final prediction, without considering the consistency and compensation. The achieved mAP performances are 79.42% on VOC-LT and 54.71% on COCO-LT, which are better than the two baselines. One possible reason is that the two branches learned from different label distributions exploit complementary information for recognizing the same label. By considering the proposed cross-branch consistency but not logit compensation, the mAP performance is improved to 81.22% on VOC-LT and 56.62% on COCO-LT, with 1.80% and 1.91% increments, respectively. Finally, we add the logit compensation to the classification loss, the mAP performance is further improved to 81.44% and 56.90%, respectively. This verifies that each component in the proposed method contributes to the mAP performance improvement.

Besides, we also show that incorporating an augmented testing (aug-test) strategy can further improve the mAP performance. In this strategy, the average of the predictions estimated from the original image and its horizontally flipped image is computed as the final prediction. Since this strategy is not widely used in the previous works, we do not consider it when comparing the performance of the proposed method against the previous methods.

Table 6.2 Ablation analysis on different components of the proposed network.

| uniform branch | | √ | | √ | √ | √ | √ |
|---|---|---|---|---|---|---|---|
| re-sampled branch | | | √ | √ | √ | √ | √ |
| logit consistency | | | | | √ | √ | √ |
| logit compensation | | | | | | √ | √ |
| aug-test | | | | | | | √ |
| VOC-LT | head | 73.14 | 71.72 | 73.98 | 75.42 | 75.68 | 76.04 |
| | medium | 83.49 | 83.58 | 84.67 | 85.50 | 85.53 | 85.92 |
| | tail | 75.41 | 79.41 | 79.56 | 82.37 | 82.69 | 83.01 |
| | total | 77.15 | 78.36 | 79.42 | 81.22 | 81.44 | 81.79 |
| COCO-LT | head | 51.61 | 48.67 | 51.81 | 54.30 | 54.13 | 54.54 |
| | medium | 57.17 | 56.87 | 58.62 | 60.27 | 60.59 | 61.10 |
| | tail | 49.21 | 50.94 | 52.06 | 53.86 | 54.47 | 54.64 |
| | total | 53.15 | 52.76 | 54.71 | 56.62 | 56.90 | 57.28 |

### Consistency Analysis

We also compare the proposed logit consistency across different training-data distributions with perturbation-based consistency and model-based consistency, as discussed in Sec. 3.3. The mAP performance from different logit consistency is reported in Table 6.3. Given a single data sampling, we add the perturbations of horizontal flipping as in Chapter 5 on the input images and feed both original and perturbed images to the ResNet50 for model learning. The consistency of the estimated logits for the original and perturbed images is considered for multi-label recognition. The perturbation-based consistency based on uniform sampling and re-balanced sampling leads to mAP performance of 78.18% and 79.39% respectively on VOC-LT, and 55.32% and 55.49% respectively on COCO-LT. While the different data distributions are merged directly, i.e. "uniform $\cup$ re-balanced", to train the network without enforcing the logit consistency, the achieved mAP performance is much lower. This is equivalent to learn the model based on another distribution that combines the uniform and re-balanced samplings.

For model-based consistency, we train the two branches with the same sampling, either the uniform sampling or the re-balanced sampling, as well as considering the consistency of logits across two branches, e.g. [172, 105]. The model-based consistency from the uniform and re-balanced samplings yields the mAP performance of 80.13% and 80.18% respectively on VOC-LT, and 55.70% and 55.44% respectively on COCO-LT. We can see that the use of the proposed consistency in our method achieves much better mAP performance than both the uses of perturbation-based and model-based consistencies on both long-tailed datasets.

Table 6.3 mAP performance by using different kinds of consistency.

| number of branches | consistency based on | sampling | VOC-LT | | | |
|---|---|---|---|---|---|---|
| | | | total | head | medium | tail |
| single | data perturbations | uniform | 78.18 | 74.09 | 83.99 | 76.90 |
| | | re-balanced | 79.39 | 73.35 | 84.71 | 79.94 |
| | N/A | uniform ∪ re-balanced | 77.85 | 72.48 | 82.68 | 78.26 |
| dual | models | uniform×2 | 80.13 | 74.71 | 85.12 | 80.46 |
| | | re-balanced×2 | 80.18 | 74.54 | 84.99 | 80.81 |
| | distributions | uniform; re-balanced | 81.22 | 75.42 | 85.50 | 82.37 |
| number of branches | consistency based on | sampling | COCO-LT | | | |
| | | | total | head | medium | tail |
| single | data perturbations | uniform | 55.32 | 52.39 | 59.60 | 52.26 |
| | | re-balanced | 55.49 | 52.01 | 59.32 | 53.50 |
| | N/A | uniform ∪ re-balanced | 53.12 | 50.14 | 57.18 | 50.38 |
| dual | models | uniform×2 | 55.70 | 52.40 | 59.28 | 53.89 |
| | | re-balanced×2 | 55.44 | 52.01 | 59.26 | 53.43 |
| | distributions | uniform; re-balanced | 56.62 | 54.30 | 60.27 | 53.86 |

Finally, we conduct an experiment to justify the proposed logit consistency against the use of the probability consistency after the sigmod normalization in the proposed network. As shown in Table 6.4, the logit consistency yields better performance than the probability consistency, by avoiding gradient vanishing as discussed in Eq. (6.8).

Table 6.4 mAP performance of the proposed network by using the logit consistency and the probability consistency, respectively.

| VOC-LT | total | head | medium | tail |
|---|---|---|---|---|
| logit | 81.44 | 75.68 | 85.53 | 82.69 |
| probability | 80.32 | 74.00 | 85.84 | 80.92 |
| COCO-LT | total | head | medium | tail |
| logit | 56.90 | 54.13 | 60.59 | 54.47 |
| probability | 56.03 | 53.11 | 59.85 | 53.55 |

### Class-wise Analysis

In Fig.6.3, we show the class-wise average precision (AP) increment made by the re-balanced branch, the branch ensemble and the proposed network, respectively, when compared to solely using the uniform branch. As shown in the top row of Fig. 6.3, compared with uniform sampling for model training, re-balanced sampling leads to AP increment on tail classes (the right portion of each curve), since it increase the sampling rate of tail-class instances. Meanwhile, it also reduces the sampling rate of some head-class images, resulting in underfitting on head-class recognition and decreased AP performance on head classes, as shown in the left portion of each increment curve in the top row of Fig. 6.3. We can see that branch ensemble can alleviate the head-class performance decrease, while keeping the AP increment in tail classes, as shown in the middle row of Fig. 6.3. The proposed method further improve the AP performance of most head, medium and tail classes by considering logit consistency between two branches and the logit compensation, as shown in the bottom row of Fig. 6.3.

To further understand the proposed logit compensation, we visualize the learned distribution parameters of Eq. (6.6) in Fig. 6.4. From the top row of Fig. 6.4, we can see that the mean values for positive and negative logit compensation are almost opposite to each other. The absolute mean value for each class largely follows a positive correlation with the sample number in this class. Since the mean values

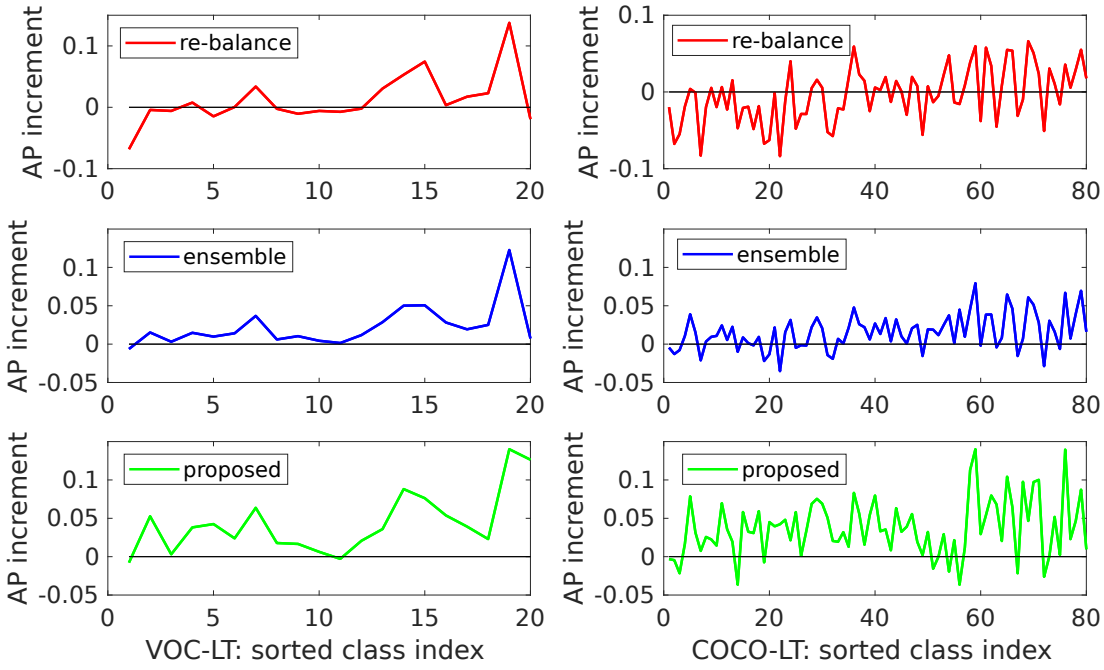


Figure 6.3 Class-wise AP increment of re-balanced branch, the branch ensemble and the proposed network over the uniform branch. Class labels are sorted from head to tail classes left-right.

for compensating logits of positive samples and negative samples are positive and negative, respectively, the absolute values of logits increases for correct predictions. This helps decrease the loss values and prevents the logit values from being away from 0 quickly. The standard deviations also approximately follow a positive correlation with the sample number in each class, as shown in the bottom row of Fig. 6.4. Besides, we can also notice that the standard deviations learned for positive logits are usually smaller than 1 and those learned for negative logits are usually larger than 1. For most classes, positive samples are usually the minority, while the negative samples are the majority. A standard deviation lower than 1 inclines to increase the classification loss from the logits, while a standard deviation greater than 1 tends to decrease the classification loss from the logits. Therefore, the loss from positive samples, along with the tail classes, are relatively emphasized to address the imbalance issue.

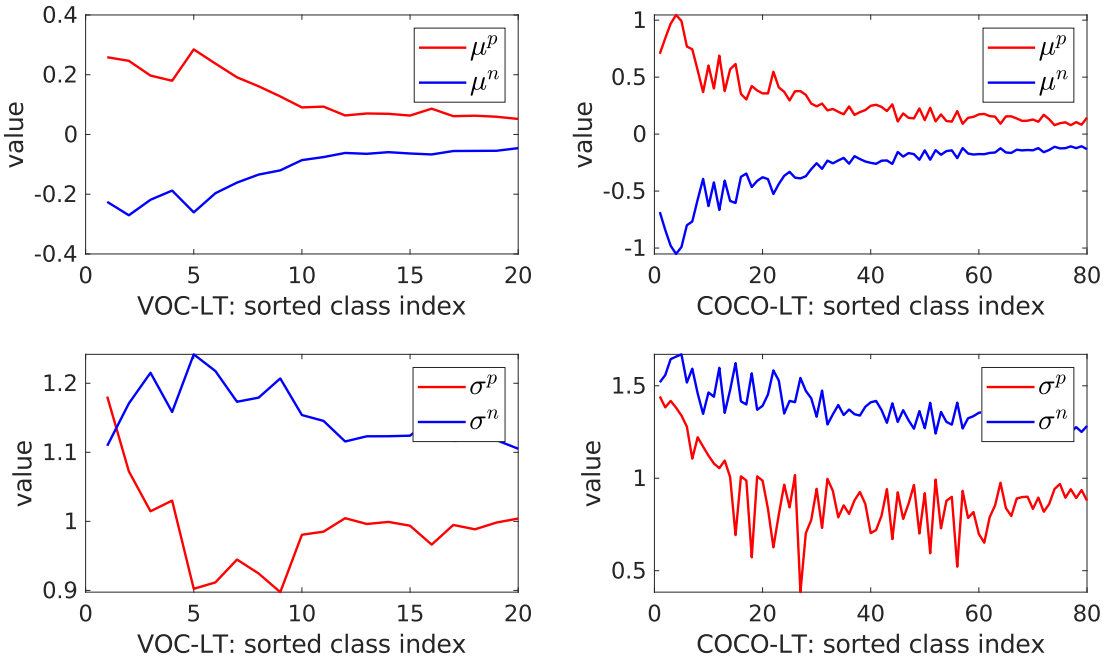


Figure 6.4 The visualization of learned logit compensation parameters for positive and negative logits, on VOC-LT and COCO-LT. Class labels are sorted from head to tail classes left-right.

### Group-wise Analysis

For all the compared methods in Table 6.1, we can notice an interesting phenomenon that mAP performance on medium classes is usually higher than those on head classes and on tail classes. The prior work [166] gives a conjecture that sample numbers of medium classes (10 to 100 samples per class) may be more suitable for the specific multi-label learning. We agree with this conjecture. With a simplified assumption that there is only one label associated to each image, a class is balanced if its number of samples is $\frac{N}{K}$. On VOC-LT, $\frac{N}{K} = \frac{1142}{20} = 57$ and on COCO-LT, $\frac{N}{K} = \frac{1909}{80} = 23.9$, both of which are in the range of $[10, 100]$ used for defining medium classes. Therefore, the sample numbers of medium classes are already more balanced than those of the head and tail classes.

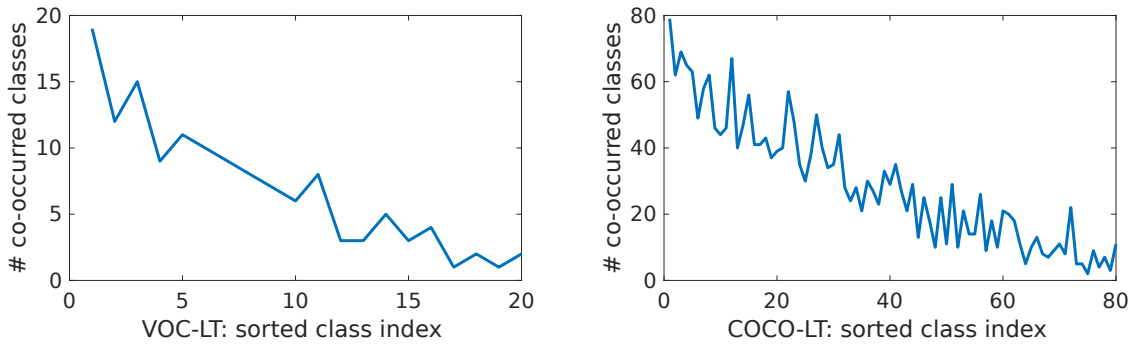


Figure 6.5 Number of co-occurred classes on the same image in term of class labels sorted from head classes to tail classes on the two datasets.

In addition, the use of re-balanced sampling, such as DB Focal, baseline-re-balanced, or the proposed method, usually leads to better performance on tail classes than on head classes, as shown in Table 6.1. One possible reason is that images with head class labels are usually associated with more classes and show more diverse and complex appearance features. As shown in Fig. 6.5, it is clear that head classes have more co-occurred classes than tail classes. In this case, without sufficient samples, the image diversity and complexity for head classes are more difficult to learn than simpler tail-class images.

### 6.3.4 Effect of Hyper-parameter $\lambda$

Besides the conventional hyper-parameters for deep network learning, the proposed method introduces one more hyper-parameter to tune, i.e. $\lambda$ in Eq. (6.3), which is end-to-end training friendly. We further conduct a set of experiments to study the effect of different configurations of $\lambda$ to the recognition performance. As shown in Fig. 6.6, when $\lambda = 0.2$, the proposed method achieves the best mAP performance of 81.49% on VOC-LT. When $\lambda = 0.1$, the proposed method achieves the best mAP performance of 56.90% on COCO-LT. An overly small $\lambda$ may not give sufficient consideration for the

consistency, while an overly large $\lambda$ may make the consistency dominate the training, leading to decreased performance on the original task of multi-label recognition.

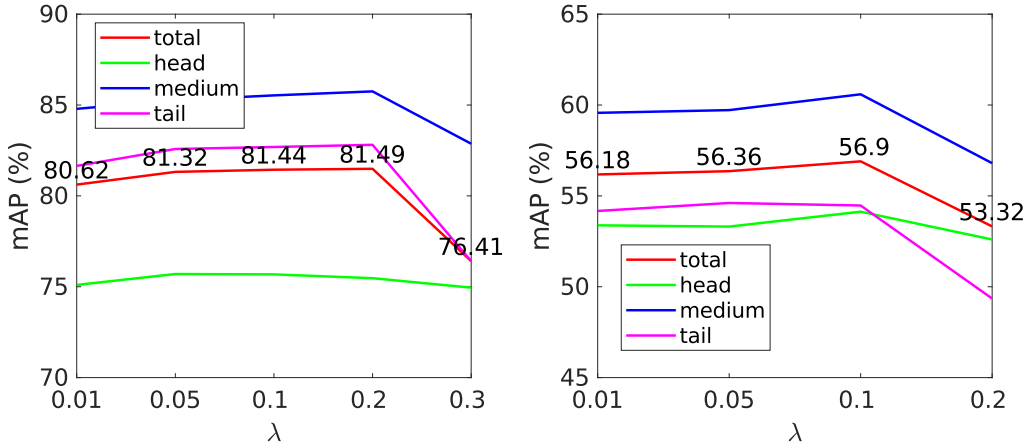


Figure 6.6 The effect of hyper-parameter $\lambda$ to the mAP performance.

## 6.4 Chapter Summary

This study tackled the task of long-tailed multi-label visual recognition by learning a model using both uniform and re-balanced samplings from the same training set. We proposed a network consisting of two branches for two samplings, respectively. Meanwhile, we incorporated the logit consistency across two branches for the same input to achieve collaborative learning. With extensive experiments on two long-tailed datasets for multi-label visual recognition, we demonstrated the effectiveness of the proposed method by achieving the new state-of-the-art performance, with significant margins over prior works.

# Chapter 7

# Conclusion and Future Work

## 7.1 Conclusion

Generally, this dissertation tackled two challenges of multi-label visual recognition, including label locality and label imbalance. To address these two issues, three novel methods were proposed in this dissertation. For label locality in human attribute recognition, the attention concentration was designed and proposed to enforce the deep network to focus on a single compact image region for recognizing each human attribute. Considering the important consistency property in computer vision, the visual attention consistency is further proposed to regularize the deep network learning, so that the estimated attention maps for human attribute recognition are more plausible. For label imbalance, this dissertation explored the collaborative learning between different samplings of the long-tailed data distribution for multi-label visual recognition, which leads to a compromise between distributions biased towards different classes and improves the recognition performance on both head and tail classes.

In the first work, we added an extra component to the deep network learning for human attribute recognition to achieve attention concentration. While minimizing the ordinary classification loss discovered image regions as evidence, in terms of attention maps, for attribute recognition, the proposed attention concentration emphasized the highlighted image regions in attention maps and suppressed the remaining image regions. Thus, minimizing the proposed attention concentration loss coincided

with minimizing classification loss when attention maps highlighted attribute-relevant regions, but confronted with minimizing classification loss when attention maps highlighted attribute-irrelevant regions. The proposed attention concentration regularized the deep network learning in an adversarial way by propelling the deep network to discover only the attribute relevant regions as the evidence for attribute recognition. The proposed method addressed an issue of part-based methods requiring accurate location of human body parts and well predefined attribute-part correlation to leverage local spatiality of human attributes. In experiments, the proposed attention concentration achieved better performance than part-based methods for human attribute recognition. Experimental results also verified that the attention concentration forced the deep network focusing attention on attribute relevant regions for human attribute recognition.

Motivated by the property of consistency in robust vision systems, this dissertation also defined and enforced the visual attention consistency to achieve consistent attribute-region relevance for human attribute recognition. Specifically, two kinds of attention consistency are defined and enforced, i.e., the attention equivariance under spatial image transforms and attention invariance between different networks. To achieve these two kinds of consistency, we designed and proposed a two-branch framework, where the ordinary classification loss for attribute learning and a new attention consistency loss were minimized simultaneously. The proposed attention consistency regularized the deep network learning by enhancing the plausibility of the attention maps for human attribute recognition. Extensive experimental results verified the effectiveness of the proposed attention consistency by achieving new state-of-the-art performance of human attribute recognition on three representative datasets. Also, experiments demonstrated that enforcing attention consistency could improve the plausibility of attention maps, which leads to recognition performance improvement.

The third work of this dissertation handled the recently arisen long-tailed issue for multi-label visual recognition. Due to label co-occurrence, existing re-balanced sampling for addressing long-tailed issue in single-label visual recognition can not achieve expected balance directly for multi-label visual recognition. Since the uniform sampling and the re-balanced sampling yielded distributions bias towards head-class and tail-class recognition, respectively, this work proposed a new two-branch network to learn from two distributions and enforced two branches to learn from each other by achieving cross-branch consistency. The proposed method regularized the deep network learning by compromising between two biased distributions and led to an effect equivalent to learning from a more balanced distribution somewhere between these two biased distributions. We conducted comprehensive experiments to verify that the proposed method improved recognition performance on both head and tail classes with substantial margins over prior works.

To summarize, this dissertation utilized two kinds of important prior knowledge in computer vision field, i.e., the visual attention mechanism and the consistency property. By addressing label locality and label imbalance of the multi-label visual recognition, we revisited these knowledge and demonstrated that deep network learning for multi-label visual recognition can be regularized with these cues.

## 7.2 Future Work

Based on the above study on multi-label visual recognition, we can outlook some of the future works. As a fundamental vision task, multi-label visual recognition is important and worth devoting effort to study. Even though several aspects of this task have been widely studied, such as label dependencies, label locality and label imbalance, there still exist a lot of issues to be addressed for multi-label visual recognition.

### 7.2.1 New Directions

**Partial Annotations for Multi-label Visual Recognition:** Existing state-of-the-art methods for multi-label visual recognition usually rely on training deep networks on large-scale image datasets, with each image annotated with multiple image labels. However, when the number of image labels associated with one image gets large, e.g., at the magnitude higher than $10^3$, it is very expensive (time and labor consumption) to annotate each image in a large-scale training set for deep network learning. Considering thousands of categories in natural scenes, this is a practical challenge for a comprehensive vision system. Based on an assumption that the number of categories in a single image is limited, label annotations for each image can be very sparse. Thus, we can explore the feasibility of using partially annotated images to learn deep networks for multi-label visual recognition. Under this circumstance, a subset of positive categories in an image is annotated as presence, while the presence of all other categories, including the negative categories and remaining positive categories, are denoted as "non-specified". Addressing this issue could marginally increase the scalability of multi-label visual recognition. For example, an extreme case is that only one of the positive categories in the image is annotated, even the image contains multiple categories. We may study the feasibility to unify the single-label visual recognition and multi-label visual recognition into the same framework, which heavily reduces the cost of annotations for multi-label visual recognition, i.e., from the number of image labels times the number of images to the number of images.

**Extremely Large-scale Multi-Label Visual Recognition with Noisy Annotations:** Also, considering the high consumption for image annotation in multi-label visual recognition, another potential solution could be using searching engines or social media for training data collection and annotation. The collected images are usually associated with certain descriptions, which can serve as the annotations for

these images. In this way, an extremely large-scale image dataset can be conveniently constructed for multi-label visual recognition. However, image descriptions can not always be trusted, leading to certain inaccurate annotations. To utilize these noisy annotations, it is necessary to design a method for deep network learning based on inaccurate annotations for multi-label visual recognition.

### 7.2.2 New Methods

There are also some aspects to explore for the original multi-label visual recognition. As discussed in this dissertation, recognizing multiple image labels can be regarded as multiple tasks of recognizing each single image label. The label co-occurring specifies the dependency between recognizing different image labels, which has been well studied in recent years. However, the independent property of each image label has not been well explored. Specifically, given different appearance complexities of different image labels, the levels of difficulties for recognizing the different image labels are usually different. During the same deep network learning for simultaneously recognizing all defined image labels, this difference has not been well curated. Several re-weighting strategies have been proposed for balancing the learning speed of different image labels based on the number of image samples for each category, without considering the intrinsic complexity of each image labels. In the experiment of this dissertation, we observed that when an image label co-occurs with more other image labels, it requires more training efforts for achieving robust recognition. This could be a cue that can be used for further improving multi-label visual recognition.

Besides, as multiple labels are learned by the same deep network, there may also exists bias competition between labels, i.e., the promising performance of recognizing one label is at the cost of low performance of recognizing some other labels, due to the different recognition difficulties and imbalanced image samples. The future work may

be still necessary to discover the optimum of compromising the recognition between different labels.

# Bibliography


[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, 2009.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[3] Naman Bansal, Chirag Agarwal, and Anh Nguyen. Sam: The sensitivity of attribution methods to hyperparameters. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8673–8683, 2020.

[4] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: A poselet-based approach to attribute classification. In *IEEE International Conference on Computer Vision*, pages 1543–1550. IEEE, 2011.

[5] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[6] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[7] Ricardo Cabral, Fernando De la Torre, Joao Paulo Costeira, and Alexandre Bernardino. Matrix completion for weakly-supervised multi-label image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):121–135, 2015.

[8] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pages 1567–1578, 2019.

[9] Kai-Yueh Chang, Tyng-Luh Liu, Hwann-Tzong Chen, and Shang-Hong Lai. Fusing generic objectness and visual saliency for salient object detection. In *IEEE International Conference on Computer Vision*, pages 914–921, 2011.

[10] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[11] Kumar Chellapilla, Sidd Puri, and Patrice Simard. High performance convolutional neural networks for document processing. In *International Workshop on Frontiers in Handwriting Recognition*. Suvisoft, 2006.

[12] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019.

[13] Ming-Ming Cheng, Niloy Mitra, Xiaolei Huang, and Shi-Min Torr, Philip Hand Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

[14] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *ACM International Conference on Image and Video Retrieval*, page 48. ACM, 2009.

[15] Dan Cireşan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. *arXiv preprint arXiv:1202.2745*, 2012.

[16] Dan Claudiu Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, and Jürgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In *International Joint Conference on Artificial Intelligence*, 2011.

[17] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pages 2990–2999, 2016.

[18] Charles E Connor, Howard E Egeth, and Steven Yantis. Visual attention: bottom-up versus top-down. *Current Biology*, 14(19):R850–R852, 2004.

[19] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

[20] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems*, pages 6967–6976, 2017.

[21] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. 2005.

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[23] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *ACM International Conference on Multimedia*, pages 789–792. ACM, 2014.

[24] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.

[25] Sander Dieleman, Jeffrey De Fauw, and Koray Kavukcuoglu. Exploiting cyclic symmetry in convolutional neural networks. *arXiv preprint arXiv:1602.02660*, 2016.

[26] Kun Duan, Devi Parikh, David Crandall, and Kristen Grauman. Discovering localized attributes for fine-grained recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3474–3481, 2012.

[27] Charles W Eriksen and James E Hoffman. Temporal and spatial characteristics of selective encoding from visual displays. *Perception & Psychophysics*, 12(2):201–204, 1972.

[28] Charles W Eriksen and James D St James. Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics*, 40(4):225–240, 1986.

[29] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

[30] Jun-Peng Fang and Min-Ling Zhang. Partial multi-label learning via credible label elicitation. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 3518–3525, 2019.

[31] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785, 2009.

[32] Pedro Felzenszwalb, Ross Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[33] Rogerio Feris, Russel Bobbitt, Lisa Brown, and Sharath Pankanti. Attribute-based people search: Lessons learnt from a practical surveillance system. In *International Conference on Multimedia Retrieval*, pages 153–160, 2014.

[34] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *IEEE International Conference on Computer Vision*, pages 3429–3437, 2017.

[35] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.

[36] Ross Girshick. Fast r-cnn. *arXiv preprint arXiv:1504.08083*, 2015.

[37] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[38] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Actions and attributes from wholes and parts. In *IEEE International Conference on Computer Vision*, pages 2470–2478, 2015.

[39] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with r* cnn. In *IEEE International Conference on Computer Vision*, pages 1080–1088, 2015.

[40] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013.

[41] Hao Guo, Xiaochuan Fan, and Song Wang. Human attribute recognition by refining attention heat map. *Pattern Recognition Letters*, 94:38–45, 2017.

[42] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 729–739, 2019.

[43] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, pages 8527–8537, 2018.

[44] Kai Han, Jianyuan Guo, Chao Zhang, and Mingjian Zhu. Attribute-aware attention model for fine-grained representation learning. In *ACM International Conference on Multimedia*, pages 2040–2048, 2018.

[45] Kai Han, Yunhe Wang, Han Shu, Chuanjian Liu, Chunjing Xu, and Chang Xu. Attribute aware pooling for pedestrian attribute recognition. *arXiv preprint arXiv:1907.11837*, 2019.

[46] Emily M Hand, Carlos Castillo, and Rama Chellappa. Doing the best we can with what we have: Multi-label balancing with selective learning for attribute prediction. In *AAAI Conference on Artificial Intelligence*. AAAI, 2018.

[47] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[48] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[50] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[51] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pages 44–51. Springer, 2011.

[52] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[53] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[54] Hexiang Hu, Guang-Tong Zhou, Zhiwei Deng, Zicheng Liao, and Greg Mori. Learning structured inference neural networks with label relations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2960–2968, 2016.

[55] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 7, 2017.

[56] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *IEEE Conference on Computer Vision and pattern recognition*, pages 5375–5384, 2016.

[57] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 3, 2017.

[58] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, 1962.

[59] David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243, 1968.

[60] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015.

[61] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7610–7619, 2020.

[62] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[63] John Jonides. Further toward a model of the mind's eye's movement. *Bulletin of the Psychonomic Society*, 21(4):247–250, 1983.

[64] Jungseock Joo, Shuo Wang, and Song-Chun Zhu. Human attribute recognition by rich appearance dictionary. In *IEEE International Conference on Computer Vision*, pages 721–728, 2013.

[65] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

[66] Andrej Karpathy et al. Cs231n convolutional neural networks for visual recognition. *Neural Networks*, 1, 2016.

[67] Jyri J Kivinen and Christopher KI Williams. Transformation equivariant boltzmann machines. In *International Conference on Artificial Neural Networks*, pages 1–9. Springer, 2011.

[68] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.

[69] Kristin Koch, Judith McLean, Ronen Segev, Michael A Freed, Michael J Berry II, Vijay Balasubramanian, and Peter Sterling. How much the eye tells the brain. *Current Biology*, 16(14):1428–1434, 2006.

[70] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[71] Neeraj Kumar, Alexander Berg, Peter Belhumeur, and Shree Nayar. Attribute and simile classifiers for face verification. In *IEEE International Conference on Computer Vision*, pages 365–372, 2009.

[72] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[73] Christoph Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, 2009.

[74] Christoph Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[75] Nilli Lavie. Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences*, 9(2):75–82, 2005.

[76] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 8(1):98–113, 1997.

[77] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178, 2006.

[78] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[79] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*, pages 396–404, 1990.

[80] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[81] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 991–999, 2015.

[82] Karel Lenc and Andrea Vedaldi. Learning covariant feature detectors. In *European Conference on Computer Vision*, pages 100–117. Springer, 2016.

[83] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *Asian Conference on Pattern Recognition*, pages 111–115. IEEE, 2015.

[84] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2018.

[85] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016.

[86] Fei-Fei Li, Andrej Karpathy, and Justin Johnson. Cs231n: Convolutional neural networks for visual recognition. *University Lecture*, 2015.

[87] Qiang Li, Maoying Qiao, Wei Bian, and Dacheng Tao. Conditional graphical lasso for multi-label image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2977–2986, 2016.

[88] Qiaozhe Li, Xin Zhao, Ran He, and Kaiqi Huang. Visual-semantic graph reasoning for pedestrian attribute recognition. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 8634–8641, 2019.

[89] Xin Li, Feipeng Zhao, and Yuhong Guo. Multi-label image classification with a probabilistic label enhancement model. In *Uncertainty in Artificial Intelligence*, volume 1, page 3, 2014.

[90] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *European Conference on Computer Vision*, pages 684–700. Springer, 2016.

[91] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[92] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[93] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95:151–161, 2019.

[94] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2970–2979, 2020.

[95] Pengze Liu, Xihui Liu, Junjie Yan, and Jing Shao. Localization guided learning for pedestrian attribute recognition. *arXiv preprint arXiv:1808.09102*, 2018.

[96] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.

[97] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *IEEE International Conference on Computer Vision*, pages 1–9, 2017.

[98] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019.

[99] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[100] Eran Malach and Shai Shalev-Shwartz. Decoupling “when to update" from “how to update". In *Advances in Neural Information Processing Systems*, pages 960–970, 2017.

[101] Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *IEEE International Conference on Computer Vision*, pages 5048–5057, 2017.

[102] Jeffrey Moran and Robert Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229(4715):782–784, 1985.

[103] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pages 4694–4703, 2019.

[104] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the International Conference on Machine Learning*, pages 807–814, 2010.

[105] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Multi-label co-regularization for semi-supervised facial action unit recognition. In *Advances in Neural Information Processing Systems*, pages 909–919, 2019.

[106] Kyoung-Su Oh and Keechul Jung. Gpu implementation of neural networks. *Pattern Recognition*, 37(6):1311–1314, 2004.

[107] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694, 2015.

[108] Seyoung Park and Song-Chun Zhu. Attributed grammars for joint estimation of human attributes, part and pose. In *IEEE International Conference on Computer Vision*, pages 2372–2380, 2015.

[109] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *European Conference on Computer Vision*, pages 135–152, 2018.

[110] Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos. Equivariance through parameter-sharing. In *International Conference on Machine Learning*, pages 2892–2901. JMLR. org, 2017.

[111] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[112] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. *arXiv preprint arXiv:1803.09050*, 2018.

[113] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[114] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[115] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*, 2018.

[116] Henry A Rowley. Neural network-based face detection. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE, 1999.

[117] Henry A Rowley, Shumeet Baluja, and Takeo Kanade. Rotation invariant neural network-based face detection. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, 1997.

[118] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[119] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning internal representations by error propagation. *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1*, 1:318–362, 1986.

[120] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):696–699, 1988.

[121] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[122] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[123] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 1163–1171, 2016.

[124] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. *arXiv preprint arXiv:1807.03903*, 2018.

[125] M Saquib Sarfraz, Arne Schumann, Yan Wang, and Rainer Stiefelhagen. Deep view-sensitive pedestrian attribute inference in an end-to-end model. *arXiv preprint arXiv:1707.06089*, 2017.

[126] Uwe Schmidt and Stefan Roth. Learning rotation-aware features: From invariant priors to equivariant descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2050–2057. IEEE, 2012.

[127] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[128] Jing Shao, Kai Kang, Chen Change Loy, and Xiaogang Wang. Deeply learned attributes for crowded scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4657–4666, 2015.

[129] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European Conference on Computer Vision*, pages 467–482. Springer, 2016.

[130] Xiaohui Shen and Ying Wu. A unified approach to salient object detection via low rank matrix recovery. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 853–860, 2012.

[131] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[132] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. *arXiv preprint arXiv:1704.02685*, 2017.

[133] Patrice Y Simard, David Steinkraus, John C Platt, et al. Best practices for convolutional neural networks applied to visual document analysis. In *IEEE International Conference on Document Analysis and Recognition*, volume 3, 2003.

[134] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[135] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[136] Liuyihan Song, Pan Pan, Kang Zhao, Hao Yang, Yiming Chen, Yingya Zhang, Yinghui Xu, and Rong Jin. Large-scale training system for 100-million classi-

fication at alibaba. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2909–2930, 2020.

[137] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *European Conference on Computer Vision*, pages 475–491. Springer, 2016.

[138] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe. Person attribute recognition with a jointly-trained holistic cnn model. In *IEEE International Conference on Computer Vision Workshops*, pages 87–95, 2015.

[139] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.

[140] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015.

[141] Zichang Tan, Yang Yang, Jun Wan, Guodong Guo, and Stan Z Li. Relation-aware pedestrian attribute recognition with graph convolutional networks. In *AAAI Conference on Artificial Intelligence*, pages 12055–12062, 2020.

[142] Zichang Tan, Yang Yang, Jun Wan, Hanyuan Hang, Guodong Guo, and Stan Z Li. Attention-based pedestrian attribute analysis. *IEEE Transactions on Image Processing*, 28(12):6126–6140, 2019.

[143] Chufeng Tang, Lu Sheng, Zhaoxiang Zhang, and Xiaolin Hu. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In *IEEE International Conference on Computer Vision*, pages 4997–5006, 2019.

[144] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, pages 1195–1204, 2017.

[145] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *Advances in Neural Information Processing Systems*, pages 844–855, 2017.

[146] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *IEEE International Conference on Computer Vision*, pages 5916–5925, 2017.

[147] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Pedestrian detection aided by deep learning semantic tasks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5079–5087, 2015.

[148] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

[149] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3), 2006.

[150] Daniel Vaquero, Rogerio Feris, Duan Tran, Lisa Brown, Arun Hampapur, and Matthew Turk. Attribute-based people search in surveillance environments. In *IEEE Workshop on Applications of Computer Vision*, pages 1–8, 2009.

[151] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[152] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

[153] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294. IEEE, 2016.

[154] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Attribute recognition by joint recurrent learning of context and correlation. In *IEEE International Conference on Computer Vision*, pages 531–540, 2017.

[155] Qi Wang, Yuan Yuan, Pingkun Yan, and Xuelong Li. Saliency detection by multiple-instance learning. *IEEE Transactions on Cybernetics*, 43(2):660–672, 2013.

[156] Xiao Wang, Shaofei Zheng, Rui Yang, Bin Luo, and Jin Tang. Pedestrian attribute recognition: A survey. *arXiv preprint arXiv:1901.07474*, 2019.

[157] Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan. Dynamic curriculum learning for imbalanced data classification. In *IEEE International Conference on Computer Vision*, pages 5017–5026, 2019.

[158] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pages 7029–7039, 2017.

[159] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020.

[160] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *IEEE International Conference on Computer Vision*, pages 464–472, 2017.

[161] Yunchao Wei, Wei Xia, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Cnn: single-label to multi-label. *arXiv preprint arXiv:1406.5726*, 2014.

[162] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *European Cconference on Computer Vision*, pages 3–19, 2018.

[163] Daniel Worrall and Gabriel Brostow. Cubenet: Equivariance to 3d rotation and translation. In *European Conference on Computer Vision*, pages 567–584, 2018.

[164] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5028–5037, 2017.

[165] Mingda Wu, Di Huang, Yuanfang Guo, and Yunhong Wang. Distraction-aware feature learning for human attribute recognition via coarse-to-fine attention mechanism. *AAAI Conference on Artificial Intelligence*, pages 12394–12401, 2020.

[166] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. *arXiv preprint arXiv:2007.09654*, 2020.

[167] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[168] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[169] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

[170] Ning Zhang, Ryan Farrell, Forrest Iandola, and Trevor Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *IEEE International Conference on Computer Vision*, pages 729–736, 2013.

[171] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1644, 2014.

[172] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.

[173] Rui-Wei Zhao, Jianguo Li, Yurong Chen, Jia-Ming Liu, Yu-Gang Jiang, and Xiangyang Xue. Regional gating neural networks for multi-label image classification. In *British Machine Vision Conference*, 2016.

[174] Xin Zhao, Liufang Sang, Guiguang Ding, Yuchen Guo, and Xiaoming Jin. Grouping attribute recognition for pedestrian with joint recurrent learning. In *International Joint Conferences on Artificial Intelligence*, pages 3177–3183, 2018.

[175] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929. IEEE, 2016.

[176] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In

*IEEE Conference on Computer Vision and Pattern Recognition*, pages 9719–9728, 2020.

[177] Zhi-Hua Zhou, Jianxin Wu, and Wei Tang. Ensembling neural networks: many could be better than all. *Artificial Intelligence*, 137(1-2):239–263, 2002.

[178] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5513–5522, 2017.

[179] Jianqing Zhu, Shengcai Liao, Zhen Lei, Dong Yi, and Stan Li. Pedestrian attribute classification in surveillance: Database and evaluation. In *IEEE International Conference on Computer Vision Workshops*, pages 331–338, 2013.

[180] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017.

# Appendix A

# List of Publications

[1] **Hao Guo**, Song Wang. "Long-Tailed Multi-Label Visual Recognition by Collaborative Training on Uniform and Re-balanced Samplings." ***CVPR 2021***.

[2] **Hao Guo**, Kang Zheng, Xiaochuan Fan, Hongkai Yu, Song Wang. "Visual Attention Consistency under Image Transforms for Multi-Label Image Classification." ***CVPR 2019***.

[3] **Hao Guo**, Brian Dolhansky, Eric Hsin, Phong Dinh, Cristian Canton, Song Wang. "Deep Poisoning: Towards Robust Image Data Sharing against Visual Disclosure." ***WACV 2021***.

[4] **Hao Guo**, Xiaochuan Fan, Song Wang. "Human Attribute Recognition by Refining Attention Heat Map." ***Pattern Recognition Letters 2017***.

[5] **Hao Guo**, Jaspreet Pandher, Michael van Tooren, Song Wang. "Process Modelling of Induction Welding for Thermoplastic Composite Materials by Neural Network." ***SAMPE 2019***.

[6] Xinyi Wu, Zhenyao Wu, **Hao Guo**, Lili Ju, Song Wang. "DANNet: A One-Stage Domain Adaptation Network for Unsupervised Nighttime Semantic Segmentation." ***CVPR 2021***.

[7] Hongkai Yu, Kang Zheng, Jianwu Fang, **Hao Guo**, Song Wang. "A New Method and Benchmark for Detecting Co-Saliency Within a Single Image". ***IEEE Transactions on Multimedia 2020***.

[8] Hongkai Yu, Kang Zheng, Jianwu Fang, **Hao Guo**, Wei Feng, Song Wang. "Co-saliency detection within a single image." ***AAAI 2018***.

[9] Hongkai Yu, Haozhou Yu, **Hao Guo**, Jeff Simmons, Qin Zou, Wei Feng, Song Wang. "Multiple human tracking in wearable camera videos with information-less intervals." ***Pattern Recognition Letters 2018***.

[10] Kang Zheng, Xiaochuan Fan, Yuewei Lin, **Hao Guo**, Hongkai Yu, Dazhou Guo, Song Wang. "Learning View-Invariant Features for Person Identification in Temporally Synchronized Videos Taken by Wearable Cameras." ***ICCV 2017***.

[11] Kang Zheng, **Hao Guo**, Xiaochuan Fan, Hongkai Yu, and Song Wang. "Identifying Same Persons from Temporally Synchronized Videos Taken by Multiple Wearable Cameras." ***CVPRW 2016***.

[12] Xiaochuan Fan, **Hao Guo**, Kang Zheng, Wei Feng, Song Wang. "Object Detection with Mask-based Feature Encoding." *arXiv preprint arXiv:1802.03934 (2018)*.